Politecnico di Torino

POLITECNICO DI TORINO

Corso di Laurea Magistrale in Engineering and Management

Tesi di Laurea Magistrale

ON-THE-JOB TRAINING AND FIRM PRODUCTIVITY IN ITALY

Evidence from Italian firms

Relatore
Prof. Luigi BENFRATELLO

Candidato
Melvin BELLASSAI
s 315815

ANNO ACCADEMICO 2024/2025

Alla mia famiglia, il mio porto sicuro.
A voi che avete sempre creduto in me, anche quando io stesso vacillavo.
A voi che avete sopportato i miei silenzi e il mio umore altalenante.
A voi che non avete mai fatto pesare i miei momenti no,
ma avete saputo esserci, sempre.

Ai miei amici, un'ancora nei momenti bui.
A voi che avete rispettato i miei tempi,
che avete accettato i miei rifiuti senza smettere di invitarmi,
che mi avete fatto ridere anche quando non ne avevo voglia.
A voi che mi avete ricordato che la vita può anche essere leggera.

Ma soprattutto a me.

A me, che ho incastrato lavoro, lezioni e scadenze
con la costante sensazione di non avere mai tempo.
A me, che ho rinunciato a ore di sonno, a weekend spensierati,
a fare tardi la sera o dare spazio a nuove persone.

A me, che non mi accontento mai,
che mi chiedo se quello che faccio sia abbastanza.
A me, che se mi fermo e guardo indietro
mi rendo conto che, alla fine,
sono fiero di me.

TABLE OF CONTENTS

Abstract

This study examines the relationship between on-the-job training and productivity, using data from INAPP's 2018 RIL questionnaire. Due to the lack of sensitive company-level data, the analysis was limited to a cross-sectional approach, making it difficult to capture the long-term effects of training. Grounded in Becker's human capital theory, the research underscores that the benefits of training typically emerge over time rather than immediately, making it inherently difficult to detect a direct impact on productivity within a single period.

Initial regression studies revealed no statistically significant association between key training variables (including total training cost, number of trained employees, percentage of trained employees, and training spend per employee) and production. Recognizing the presence of heteroscedasticity and influential outliers, robust regression approaches were used to modify the estimations. This method produced more accurate results, demonstrating that the total number of trained staff had a statistically significant beneficial effect on production. In contrast, the percentage of trained personnel showed a negative relationship, indicating that while training a larger workforce may be advantageous to some extent, excessive investment in training does not always result in equal productivity increases.

Sectoral analysis revealed that the impact of training varies greatly by industry. In manufacturing, productivity is mostly determined by the quantity of trained people, while training fees and per-employee costs have no discernible impact. In some industries, such as construction and entertainment, training costs appear to be negative, most likely due to inefficiencies or operational disruptions. In others, on-the-job training has no impact on productivity.

Despite these findings, the cross-sectional nature of the data precludes causal interpretation. The study underscores the notion that the benefits of training on productivity are multifaceted and frequently evolve over time.

Keywords: on-the-job training, productivity, human capital, robust regression, INAPP data, RIL questionnaire, sectoral analysis.

1. INTRODUCTION

1.1 RESEARCH CONTEXT

The notion of human capital has been a prevailing issue in managerial and economic writings[1] for the past several decades. Ongoing employee learning is now a key component of corporate productivity[2] improvement, as well as maintaining companies competitive in an increasingly rapid, technologically sophisticated environment. Although actual productivity gains from training are sometimes difficult to quantify and depend on several factors, such as the type of training, level of employee skills, and applicable economic conditions, the new digitalization and changing labor market needs call for more focus on training as a mechanism of innovation and adaptation.

Perhaps the most contentious issue associated with workplace training is whether or not it is of any benefit to the individuals and to the companies that have sponsored such training. Thus, training increases human capital for the worker. This in turn increases employability and the probability of promotion and may lead to greater prosperity. The argument rests on the economic theory of human capital hypothesis (Becker, 1964). Within the theory in which companies invest in training, there will be a contingent improvement in the productivity of that company, since one expects a more productive workforce to yield yet greater performance, innovation, and thus returns.

1.2 THESIS PURPOSE

The purpose of this thesis is to analyze how useful corporate training is for the workers and the impact it has on business productivity, with specific reference to the Italian context. The paper will consider both the theory of major economic views on human capital and training, particularly Becker and Dearden, and empirical research conducted in some countries. Finally, data analysis from INAPP[3] will be presented in order to check whether the empirical evidence collected in Italy is in line with what has been observed in the international literature.

For those reasons, the research questions that will be guiding this analysis are the following:

1. How does corporate training impact employee productivity and, as a result, business performance?
2. Do the Italian empirical data confirm or contradict the main economic theories on training and productivity?
3. Do firms do better if they have more graduated employees than not graduated ones?

[1] See Schultz, T. W. (1961) and Becker, G. S. (1964)
[2] See Psacharopoulos, G., & Patrinos*, H. A. (2004)
[3] Istituto Nazionale per l'Analisi delle Politiche Pubbliche (National Institute for the Analysis of Public Policies)

In order to assess the goal, the starting point will be the human capital theory as propounded by Gary Becker in 1964, and the following reviewed reprints. According to Becker, training equates to an investment that raises the level of skills and productivity of the worker, and thus it benefits both the worker and the company. To further support this perspective, the work of Dearden et al. (2006) will also be reviewed; it has provided evidence of a positive correlation between investment in training and business performance. Not all sources lead to similar conclusions, as some studies underscore the existence of contextual factors, which may restrain the effectiveness of training and hence decrease its influence on productivity.

The empirical analysis of this thesis focuses on the data from INAPP concerning corporate training in Italy. By contrast to many other works, the statistics provided do not present data about rewards for the employee; direct empirical analysis of a connection between the level of training and the corresponding level of compensation is therefore not possible. For the sake of completeness, the results of some studies that have dealt with the topic will be reported.
Additionally, the thesis will briefly investigate whether organizations with a higher composition of university graduates perform better. However, this component will not be examined in depth, but rather as a thought-provoking issue to prompt more thinking.
Nevertheless, it will be possible to check whether the trend of the training distribution at the various company levels follows the theoretical forecasted one and to compare the results with those that have emerged in other international studies.

1.3 STRUCTURE OF THE THESIS

Methodologically, the thesis will follow a structure that combines theoretical analysis with an empirical verification. First, the main economic theories of training and human capital will be presented, followed by a review of existing empirical studies. In the second part, an analysis of the INAPP data will be presented to assess the relationship between training and productivity in the Italian context. Finally, the results obtained will be discussed, comparing them with the literature and analyzing their implications for Italian companies.

Limitations of the research.

This thesis presents some limitations that must be taken into consideration when interpreting the results. First, the data provided by INAPP are not detailed enough to enable direct analysis of the impacts of training on wages and individual career paths for employees and, therefore, the estimate

of eventual economic returns for the workers. Secondly, the analysis is based on observational data, which implies the impossibility of establishing a certain causal link between training and business productivity. In addition, the Italian context shows specific economic, cultural, and regulatory features that may affect the results and reduce the possibility of generalizing the same to other countries. Lastly, the thesis does not consider other variables that may influence productivity, such as the company climate, personnel management policies and the level of technological innovation. In fact, it would be very difficult to take into account, as stated by Angrist and Pischke, (2009)[4].

This research, therefore, adds some empirical evidence to the debate about the importance of corporate training, though now specifically referring to the Italian context. If the results confirm the validity of Dearden's theory, this suggests to companies and policymakers the need to further incentivize investments in training in order to improve the competitiveness of the national economic system.

[4] See Angrist, J.D., & Pischke, J.S. (2009)

2. LITERATURE REVIEW

Human capital is one of the milestones of modern economic theory, representing one of the basic under-pinners in understanding the relationship between education and training with economic growth. In this regard, *human capital* defines the set of knowledge, skills, abilities, and experiences that an individual has acquired and that enhances his productivity within the labor market. The concept has gained wide relevance in economic analysis because it recognizes the possibility that an investment in people's education and training can be fruitful both for the individual and the community. The concept of human capital arises from the need to explain income differences between individuals and nations and the role of education and training in economic growth (Psacharopoulos, G. et Al., 2004). In the currently more knowledge-based and innovative-driven economy, human capital is considered a crucial productive factor alongside physical capital. Its value has been increasingly realized not only in the economic sphere but also in the political and social spheres, with numerous policies aimed at promoting training and skills development.

The notion of human capital has always been the source of debates and studies from classical theorists such as Adam Smith and Alfred Marshall, through to contemporary figures like Gary Becker and Theodore Schultz. Indeed, only with the second half of the 20th century did the formalization of human capital theory shed light on the process whereby investment in education and training affects individuals' and groups' productivity and has related consequences for the labor market and competitiveness.

The following paragraphs discuss in fuller detail the theory of human capital by Gary Becker and the models proposed by Dearden et Al., which has given one of the most instructive analyses in the area concerned.

2.1 BECKER'S HUMAN CAPITAL THEORY

Becker's theory of human capital addresses numerous topics, ranging from education to training, encompassing the influence of family, school, and the environment in which one lives. For the sake of simplicity and conciseness, this thesis will only cover the theories regarding the effects of education and on-the-job training on human capital and the respective companies.

Education and training are the most important investments in human capital. A large body of research has firmly established that the higher the educational attainment of individuals, the higher their incomes. Studies in the United States and replicated in over one hundred countries with very different cultural and economic systems, show convincingly that high school and college education greatly

raise individual earnings. This is true even when the direct and indirect costs of schooling are accounted for, and even after controlling for family background and other factors that may be correlated with higher education. The general pattern is that those with more education tend to earn considerably above-average incomes.

Of course, formal schooling is not the only way that people learn and develop skills. A great deal of learning occurs outside of formal educational settings, and especially on the job. University-educated workers frequently must be upgraded in the job market before it is productive and useful. Sometimes this on-the-job training (OJT) may involve very short periods – for example, two weeks of job orientation for people working in certain jobs. Frequently, it could take many months or even up to three-four years in specialized jobs like being an engineer or doctor. Existing evidence on such OJT does suggest that there is considerable earning growth as job experience accumulates.

Reciprocal investment in OJT ties workers and employers together (Schultz, T.W., 1961), which explains why job turnover rates for unskilled workers are higher than those of skilled workers. In addition to human skill acquisition, work attitudes are a main driver of work behaviours according to research conducted by Harrison, Newman, and Roth (2006). Their meta-analysis demonstrates that a general, positive job attitude encourages a disposition to contribute positive inputs to one's work role, thus strengthening the connection between effective on-the-job training, employee motivation and productivity.

The relationship between training and retention is further nuanced when one considers generational differences among employees. According to research by D'Amato and Herzfeldt (2008), younger generations, in particular Early and Late Xers[5], have demonstrated a greater learning orientation that can impact their intention to remain with an organization. This means that creating an environment for ongoing learning opportunities yields not just a higher productivity but also better retention — especially for millennial managers. Comparisons across countries, such as that between Japan and the United States, indicate that greater reliance on OJT is one of the reasons for the lower rates of job mobility in Japan with respect to United States.

Although it is comparatively easy to quantify the monetary return to human capital, research has also made a beginning in assessing its non-monetary return. Many studies indicate that education is associated with improved health outcomes, reduced smoking rates, increased civic engagement (such as voting), better knowledge of family planning, and a greater appreciation for cultural pursuits. Groundbreaking research by Bob Michael (1972), employing economic theory, has attempted to

[5] Early Boomers (born between 1946 and 1951), Late Boomers (born between 1952 and 1959), Early Xers (born between 1960 and 1970), and Late Xers (born between 1971 and 1980).

quantify some of these non-monetary benefits of education. His estimates, along with those of several other researchers, indicate that such benefits are significant, but for most people they are probably less significant than the monetary rewards.

Focusing on OJT, the goal is to formalize the recognition of the job itself on productivity. The premise is that many workers improve their skills and refine old ones through workplace learning, resulting in higher productivity. This future productivity growth is not costless; rather, it requires an investment in training.

These costs represent an opportunity cost: resources used to develop future output could otherwise have been deployed for current or increased production. Recent data has supported this claim by showing how training during employment increases the skills of workers and improves productivity at the firm beyond the value of increased wages. For example, a firm's productivity increases between 1.7% and 3.2% when the number of trained workers increases by 10%, while the firm's wages bill only increases between 1.0% and 1.7% (Konings & Vanormelingen, 2015).

Let's consider a firm hiring employees for a specific period (plausibly a very short one), assuming for now perfectly competitive labor and product markets. Absent any on-the-job training, the firm faces a fixed wage rate, outside of its control. A profit-maximizing firm, under these conditions, reaches equilibrium when the marginal product of labor equals the wage rate, effectively equating marginal revenue and marginal cost; in symbols:

$$MP = W \tag{1}$$

where W and MP represent, respectively, wages and marginal product (or receipts). A more comprehensive equilibrium state can be represented as:

$$MP_t = W_t \tag{2}$$

where *t* denotes the *t*-th period. The equilibrium in each period is determined solely by economic activity within that specific timeframe.

$$\sum_{t=0}^{n-1} \frac{R_t}{(1+i)^{t+1}} = \sum_{t=0}^{n-1} \frac{E_t}{(1+i)^{t+1}} \tag{3}$$

where:

- n represents the number of timeframes,
- R_t and E_t are functions of all the other receipts and expenditures.

The generalized equilibrium condition replacing the simpler one in equation (2) is that the present value of the stream of marginal products equals the present value of the wage stream. Whereas marginal product equating wages in every period implies this present value equality, the reverse is not necessarily true. A balance in present values does not require period-by-period equality between marginal product and wages.

If training were to occur only in the first period, initial period expenditures include wages plus the cost of training. In all subsequent periods, expenditures include only wages while revenues in all periods are determined by marginal productivity. Equation (3) becomes:

$$MP_0 + \sum_{t=1}^{n-1} \frac{MP_t}{(1+i)^t} = W_0 + k + \sum_{t=1}^{n-1} \frac{W_t}{(1+i)^t} \tag{4}$$

where *k* represents the disbursement on training.

So, it is possible to define a new term

$$G = \sum_{t=1}^{n-1} \frac{MP_t - W_t}{(1+i)^t} \tag{5}$$

representing the difference between future revenue and future expenditure, so it quantifies the firm's return on its training investment. The equation (4) can be rewritten as

$$MP_0 + G = W_0 + k\,. \tag{6}$$

While *k* represents the actual expenditure on training, it doesn't fully capture the total cost. It omits the opportunity cost of the trainee's time, which could otherwise have been spent generating current output. That opportunity cost is equal to the difference between what could have been produced, $MP_0{}'$, and what is currently produced, MP_0. Labelling *C* as the overall sum of opportunity costs and disbursement on training, equation (6) changes in

$$MP_0{}' + G = W_0 + C. \tag{7}$$

Consequently, the difference between G and C reflects the net return on training, considering both its benefits and costs. Equation (7) reveals that the marginal product in the initial period equals wages only when the return on training matches its cost (G = C). If the return is less than the cost (G < C), the marginal product will be less than wages. Conversely, if the return exceeds the cost (G > C), the marginal product will exceed wages.

At this point, Becker introduces two basic concepts in the context of on-the-job training, namely general and specific training. To complete these concepts are mentioned here in this thesis, but for brevity, they will not be treated to the same extent as in the original text.

General training

General training provides workers with transferable skills, increasing their productivity across multiple firms. In competitive labor markets, this leads to higher wages and marginal products.

Human capital theory served as the basis for the models used to explain the relationship between training and productivity, focusing on the fact that general training not only enables workers to be more productive proponents of the company but to all industries (De Grip and Sauermann, 2013). In contrast, general training increases productivity throughout the employee's career, which weakens the employer's incentive to invest in skill development because they cannot compel their employees to remain employed and benefit from the increased productivity.

However, firms cannot capture returns from *perfect* general training because wages adjust accordingly. Therefore, firms only offer general training if the cost is borne by the trainees themselves, who ultimately benefit from the resulting wage increases. Starting from equation (7), and since marginal products and wages are raised by the same amount, MP_t has to be equal to W_t for all periods. Therefore,

$$G = \sum_{t=1}^{n-1} \frac{MP_t - W_t}{(1+i)^t} = 0\,, \tag{8}$$

and so, equation (7) is reduced to

$$MP_0' = W_0 + C\,. \tag{9}$$

In terms of actual marginal product

$$MP_0 = W_0 + k\,. \tag{10}$$

Trainees' wages will be lower than their potential marginal product by the full cost of the general training. Essentially, employees pay for general training through reduced current earnings. Equation (10) has numerous further implications, briefly:

- Wage Depression during training: employees really pay for general OJT by accepting lower compensation during the training term. To put it another way, the gap between their potential production and actual pay represents the cost of the training they receive.
- Intermingling of capital and income: while income (potential productivity) and capital (training expenditures) are clearly separated in the context of material products accounting, they are both included in training earnings. This is because, in contrast to the depreciation of physical capital, investments in human capital are written off immediately rather than over time.
- Low present earnings, high future earnings: although trainees may have exceptionally low, or even negative, current "incomes," their long-term earnings are anticipated to be higher as a result of the skills they have learned. Correlations between incomes and current spending may become distorted as a result.
- Human capital depreciation: it is a phenomenon that is frequently disregarded. In the case of OJT, the lower pay during the training phase reflects this depreciation. Although the timing of this "write-off" is different from that of physical capital, the idea is the same.
- Steeper age-earnings profiles: training causes concave and steeper age-earnings curves. This is because training (the investment phase) results in lesser earnings, while the return phase results in larger earnings.
- Cost of foregone earnings: like direct expenses, foregone earnings while training represent a considerable cost. All training expenses—direct or indirect—represent lost revenue.
- Firms don't pay for general training: because they can't profit from general training in competitive markets, businesses won't pay for it. Since they will benefit from improved skills and future earnings, trainees will pay for it with lower compensation.
- Property rights in skills: just like innovations, which need patents, skills belong to the individual by nature. This "property right" explains why businesses can transfer training costs to employees and encourages them to invest in training by accepting lower compensation. This sets training investment apart from research and development investment.

Specific training

Specific training increases productivity more in the firm that provides it. Fully specific training has no value outside the original firm, much on-the-job training falls between pure general and pure specific.

A good example of specific training is military training: although some skills are transferable to civilian life, other skills, such as those of astronauts or fighter pilots, have limited applicability. Onboarding also falls into the category of specific training because the knowledge acquired has greater benefit to the company than to other external firms. The actual costs of hiring, including recruitment fees, interviews, and background checks, are not considered training; however, they are human capital investments because it is lost if employees leave.

Companies also invest in assessing employees' abilities through testing, job rotation, and other methods. These investments are specific if the acquired knowledge remains proprietary, enhancing efficiency only within the firm. The impact of such investments on the labor market depends on competition: in highly competitive markets, the risk of losing trained employees reduces incentives for specific training.

If all training were pure specific, other firms' employees' wages would not reflect previous training, and the firms themselves would bear its full cost by reaping higher productivity. If there is competition, the present value of the returns from training must equal costs.

These propositions can be stated formally: according to equations (5) and (7), a firm that provides training in competitive markets finds its equilibrium when

$$MP_0' + G\left[= \sum_{t=1}^{n-1} \frac{MP_t - W_t}{(1+i)^t}\right] = W_0 + C\ , \tag{11}$$

in which C is the cost of training given only in the first period, MP_0' is the opportunity marginal product of trainees, W_0 is the wage paid to trainees, and W_t and MP_t are the wage and the marginal product in timeframe *t,* respectively.

Keeping in mind the definition on specific training, W would always equal the wage that could be given by someone else, $MP_t - W_t$ would be the full return in timeframe t from beginning-training, and G would be the present value of these returns. Since MP_0' represents the marginal product elsewhere and W_0 the trainees' wage elsewhere,

$$MP_0' = W_0\ . \tag{12}$$

As a consequence,

$$G = C \tag{13}$$

or, in full equilibrium, the return from training equals costs.

Before assuming that the normal equality between wages and marginal product applies to entirely particular training, two important considerations need to be made. First, prospective or opportunity marginal product (rather than actual marginal product, which may be lower because of training time) is the focus of the first wage-marginal product equality. Second, wages will eventually fall even if they initially match the marginal product. The reason for this is that the company uses the gap between future marginal products and wages to recover its specialized training investment. According to this structure, businesses pay all expenses and receive all benefits. An alternative viewpoint, on the other hand, is just as tenable: employees could pay for certain training with lower starting salaries and then receive compensation in later periods that matches their marginal product.

In terms of equation (11), W_t would equal MP_t, G would equal zero, and W_0 would equal $MP_0' - C$, just as with general training. Is it more likely that firms, rather than workers, both invest in and benefit from training?

Let's consider this: if a firm invests in a worker's specific training and that worker then quits, the firm loses part of its investment. Similarly, a worker who is dismissed after paying for specific training forfeits their expected return. Therefore, the incentive for both firms and workers to invest in specific training is strongly tied to the probability of continued employment.

Since labor turnover is usually not included in conventional economic theory, introducing it at this point may appear to be an unnatural intervention. According to standard competitive firm analysis, turnover is negligible since wages are assumed to be equal to marginal product across businesses. Because departing employees may easily find equal roles elsewhere and employers can replace them without impacting profitability, it makes no difference if a company's personnel is steady or continually shifting. Turnover is essentially ignored by traditional theory since, in its oversimplified framework, it is deemed inconsequential. Since turnover adds expenses for both businesses and employees, it becomes a crucial consideration when thinking about specialized training. A company loses money if it spends money on specialized training for an employee who later leaves. A replacement can be hired at the same market wage, but it will probably be less productive at first and will cost more to catch up with the former employee. As a result, the company loses money since the skilled employee leaves. In a similar vein, an employee who pays for specialized training and is subsequently let go loses out since they are unable to find a job that offers the same benefits

elsewhere. As a result, adding turnover to the specific training analysis is a necessary byproduct of their interdependence rather than a discretionary choice.

Businesses that invest in specialized training can account for employee turnover by aiming for a high enough return from their current workforce to offset the losses from departing workers. However, this "success-only" return computation will overestimate the average return on all training investments.

Businesses may proactively lower turnover by providing post-training compensation above the market rate, as opposed to merely making up for training losses from employee turnover by obtaining higher returns from those who remain. To balance supply and demand for training, it is sense to share training expenses and benefits with staff members. According to this last scenario, businesses and employees share both the expense and the reward. The exact sharing plan is determined by a number of variables, including the correlation between quit rates and earnings, layoff rates and profits, and other characteristics including liquidity preferences, risk tolerance, and capital access. Training increases productivity and, hence, pays at other companies if it is not entirely particular.

One way to think of this kind of instruction is as a blend of entirely generic and entirely specific elements. The general component increases with the impact on wages at other enterprises in comparison to the training firm. The percentage of costs borne by businesses is inversely proportional to the size of the general component, or positively related to the specificity of the training, because firms only support a fraction of the specific training costs and none of the general training costs.

Those conclusions can be stated formally as before: if G represents the training's present value collected by firms, the fundamental equation is

$$MP' + G = W + C. \tag{14}$$

Defining G' the employees' return collection, the sum of G and G' would be the total return, called G''. In full equilibrium $G'' = C$. Be a represent the fraction of firms' total collected return. Since $G = aG''$ and $G'' = C$, equation (14) becomes

$$MP' + aC = W + C\ , \tag{14a}$$

or

$$W = MP' - (1 - a)C\ . \tag{14b}$$

Employees bear the same proportion $(1 - a)$ of training costs as they receive in returns, generalizing the previous findings. Specifically, if training is entirely general $(a = 0)$, this relationship simplifies the earlier equation (9). Instead, if firms collected all the return from training $(a = 1)$ equation (14) simplifies to $MP_0' = W_0$. In all other cases $(0 < a < 1)$, none of the previous equations is satisfactory, highlighting the necessity of the more general equation (14).

2.2 THE CONTRIBUTION OF DEARDEN ET AL.

The importance of corporate training in the modern economic and organizational context has been the subject of numerous academic studies. The relationship between training expenditures and increased productivity is a major topic in managerial and economic research, although it is still unclear how training influences both individual and corporate performance.

The primary hypotheses that explain the econometrical connection between production and training will be examined in this chapter, with a special emphasis on the idea put forth by Dearden et al. (2006)[6]. Supporting these hypotheses, D'Amato and Herzfeldt (2008) reported that learning orientation and leadership development intentions are fundamental in achieving talent retention, in particular in special regard of different generation cohorts. Their study shows increased productivity from training, but whether it works on retention is highly dependent on the generation of the employees and their commitment to the organization.

Dearden's approach, which provides an analytical framework that explicitly connects training investments to business performance, has marked a turning point in the study of how training affects the value added produced by businesses.

The chapter will begin with an analysis of Dearden's theoretical model and conclude with a review of empirical research that has emerged from it. The primary contributions of global literature will be examined, emphasizing the findings that support the positive relationship between productivity and training as well as the important problems and potential restrictions that surfaced in later research.

These results confirm that although investments in human capital are beneficial, their effects are moderated by industry-specific elements and firm approaches. To do this requires a more refined approach to how training is designed and implemented, which also matches employee considerations with the organization (Konings & Vanormelingen, 2015; Singh & Mohanty, 2010).

To provide the foundation for the empirical analysis that will be built in the upcoming chapters, this part aims to present a clear and comprehensive picture of the body of knowledge currently available on the subject.

Furthermore, as Memon, Salleh, and Baharom (2016) showed, training isn't just about making people work better, it actually affects how much they love their jobs and whether they want stick around or not. This is like a two-for-one deal for companies because good training can give them both smarter employees and a more loyal team. But here's the thing, as Konings and Vanormelingen (2015) pointed

[6] See Dearden, L., Reed, H., & Van Reenen, J. (2006)

out, there's this productivity-wage gap happening. Basically, after workers get trained, companies might not give them the full financial pat on the back for their extra effort. This makes companies cheap, so they don't train people as much as they should. And that's because they're not thinking about all the extra good stuff training can bring, like how it helps other companies down the line too. Because of these messed up job market things, sometimes the government has to step in with some training cash to make sure everyone gets the skills they need and businesses don't skimp out on investing in their people.

Dearden's theoretical model.

There was, at that time, broad academic agreement that the UK must boost work-related training in order to improve long-term economic performance and close the "skills gap". Although there is a lot of literature on human capital investment and considerable policy interest, few studies have specifically looked at how training affects productivity; instead, they have focused on wage effects. This gap is filled by the study of Dearden et al. (2006), which offers an analysis based on a panel dataset spanning 14 years.

In traditional economic models, salaries are assumed to reflect the marginal productivity of labor in a market with perfect competition. This link, however, may be skewed by models like Becker's (1964) and labor market frictions, which indicate that training might boost productivity without always resulting in a corresponding rise in wages. Though they frequently include methodological flaws including single-point training measures and trouble identifying particular effects, prior research has shown a favorable link between training and company output.

By using a 14-years panel dataset, Dearden et al.'s study mitigates attenuation bias caused by measurement errors by treating corporate training as an endogenous variable using GMM estimators. This is the brilliant novelty of the study. Furthermore, it is possible to evaluate whether trained workers receive compensation commensurate with their marginal output by evaluating the impact of training on salaries and productivity. Dearden et al. found that a 1% increase in training is linked to a 0.6% boost in productivity and a 0.3% increase in pay, according to the research, which was mostly done at the industry level.

It's not as simple as "training equals more productivity". Studies show that how much training actually helps employees depends a lot on the industry they're in: what works for a tech company might be completely different from what a manufacturing plant needs. As researchers Singh and Mohanty pointed out back in 2010, different industries have different needs. To really get the most out of training, companies need to create programs that are specifically designed for their sector.

This suggests that studies that just look at wages understate the entire advantages of training. Dearden et al.'s economic model has been analyzed in this chapter, looking at its methodological ramifications and empirical results in light of previous research.

The model of training and productivity.

To illustrate the methodology, let's assume the possibility of modelling a representative plant within an industry using a Cobb-Douglas production function expressed in terms of value added.

$$Q = AL^{\alpha}K^{\beta} , \tag{15}$$

where Q stands for value added, L for effective labor input (which takes into account both quantity and quality), K for capital, and A for a Hicks neutral efficiency parameter. Assume that trained workers are more productive than untrained ones, so that effective labor input can be written as

$$L = N^{U} + \gamma N^{T} , \tag{16}$$

in which N^{U} and N^{T} stand for the number of untrained and trained workers, respectively, and γ is a parameter which, if trained workers are more productive than untrained ones, is greater than 1.

The total number of workers, namely N, is equal to the sum of N^{U} and N^{T}. Considering the portion of trained workers within an industry, denoting it $TRAIN = N^{T}/N$, and substituting equation (16) in equation (15) gives:

$$Q = A[1 + (\gamma - 1)TRAIN]^{\alpha}N^{\alpha}K^{\beta} . \tag{17}$$

Taking natural logarithms

$$lnQ = lnA + \alpha \ln[1 + (\gamma - 1)TRAIN] + \alpha lnN + \beta lnK . \tag{18}$$

This could be estimated by non-linear least squares. If $[(\gamma - 1)TRAIN]$ is quite "small", it is possible to use the approximation $ln(1 + x)=x$ and rewrite the production function as

$$lnQ = lnA + \alpha(\gamma - 1)TRAIN + \alpha lnN + \beta lnK . \tag{19}$$

Supposing that the industry manifests constant returns to scale (so, $\alpha + \beta = 1$), then equation (19) can be rewritten in terms of labor productivity as

$$\ln\left(\frac{Q}{N}\right) = lnA + (1-\beta)(\gamma-1)TRAIN + \beta\ln\left(\frac{K}{N}\right). \quad (20)$$

Where the productivity of trained and untrained workers is equivalent (γ = 1), the coefficient of TRAIN will be zero.

This method is readily extendable to cover more than a single category of heterogeneous workers in the labor quality index. By indexing each type of labor with *k*, equation (18) can be extended as follows:

$$\ln(Q) = lnA + \alpha ln\left\{1 + \sum_k \left[(\gamma_k - 1)\left(\frac{N_k}{N}\right)\right]\right\} + \alpha lnN + \beta lnK\,. \quad (21)$$

The empirical model permits a multi-dimensional characterization of labor quality in terms of education, occupation, age, tenure, and gender. To account for all the other variables affecting productivity (measured by A), the model also has controls for differential hours, worker turnover rates, innovation (measured by research and development expenditure), regional mix, and the proportion of small firms.

Letting *X* represent the aforementioned factors, and subject to the assumptions of constant returns to scale and a log-linear functional form, the production function is given by:

$$\ln\left(\frac{Q}{N}\right) = (1-\beta)\sum_k \left[(\gamma_k - 1)\left(\frac{N_k}{N}\right)\right] + \beta\ln\left(\frac{K}{N}\right) + \delta' X\,. \quad (22)$$

The wage equation estimated parallels the productivity equation in (22). Furthermore, the wage equation functions as a descriptive regression rather than a structural representation.

According to the notion of marginal productivity, relative wages in competitive labor markets correspond to relative marginal productivities. Because businesses would only hire trained people, a situation in which the relative productivity of trained workers (γ) surpasses their relative pay is unsustainable.

Think of a typical plant in an industry, and the associated wage cost is represented by *W*. Let assume a simplified labor market with two worker types: trained workers earning an average pay (w^T) and untrained workers earning an average wage (w^{NT}). Relative wages are $\lambda = w^T/w^{NT}$. By definition

$$W = w^{NT}(N - N^T) + \lambda w^{NT} N^T = w^{NT}[N + (\lambda - 1)N^T]\,. \tag{23}$$

In logarithms, considering $a = \ln(w^{NT})$, the average wage *w* is

$$\ln w = \ln\left(\frac{W}{N}\right) = a + \ln[1 + (\lambda - 1)TRAIN]\,. \tag{24}$$

Equation (24) can be estimated to determine the relative wage mark-up for training, λ. This estimate can then be compared with the estimated relative productivity effect of training, γ.

The wage equation, like the productivity equation, will include a wide range of explanatory variables, including various aspects of labor quality, capital inputs, and other pertinent variables. Thus, the following is the empirical wage equation that needs to be estimated:

$$\ln w = a + \sum_k \left[(\lambda_k - 1)\left(\frac{N_k}{N}\right)\right] + \beta^w \ln K + \delta^{w\prime} X\,. \tag{25}$$

The econometric model.

Let's start from the basic equation

$$y_{it} = \theta x_{it} + u_{it}\,, \tag{26}$$

Where *y* is *Q/N* and *x* is a vector of (suspected endogenous) variables including training. Subscript *i* stands for the representative firm in an industry, *t* is the timeframe and θ is the parameter of interest. Assume the stochastic error term, u_{it}, takes the form

$$u_{it} = \eta_i + \tau_t + \omega_{it}\,, \tag{27a}$$

where

$$\omega_{it} = \rho\omega_{it-1} + \nu_{it} \tag{27b}$$

The model's error term is decomposed into four terms:

- η_i , capturing fixed effects for each individual;

- τ_t , macroeconomic shocks identified by a set of time dummy variables;
- ω_{it} , a term that follows a first-order autoregressive AR(1) process with parameter 'ρ'. The existence of this term is accounted for by errors of measurement or slow technological change dynamics, and these impacts are added to the error term.
- ν_{it} , an i.i.d. error term with zero mean, which satisfies the normal assumptions of a random error;

Substituting equations (27a) and (27b) in equation (26), gives the dynamic equation:

$$y_{it} = \pi_1 y_{it-1} + \pi_2 x_{it} + \pi_3 x_{it-1} + \eta_i^* + \tau_t^* + \nu_{it} . \quad (28)$$

Note that the common factor restriction is $\pi_1\pi_2 = -\pi_3$, $\tau_t^* = \tau_t - \rho\tau_{t-1}$ and $\eta_i^* = (1-\rho)\eta_i$.

How best to estimate equation (28)? Assuming training is strictly exogenous and ρ=0 (there are no dynamics), the only issue with OLS on equation (26) is individual effects (η_i). If η_i and x_{it} are uncorrelated, then the random-effects estimator is unbiased and efficient. If correlated but exogenous, the random-effects estimator is biased, but not within-groups. When training is endogenous, instrumental variables are needed. Lacking natural experiments, it is necessary to use moment conditions to build a GMM estimator for equation (28).

A standard approach involves applying first differencing to equation (28) in order to eliminate the fixed effects:

$$\Delta y_{it} = \pi_1 \Delta y_{it-1} + \pi_2 \Delta x_{it} + \pi_3 \Delta x_{it-1} + \Delta\tau_t^* + \Delta\nu_{it} . \quad (29)$$

At this point Dearden et al. relie on Arellano and Bond, 1991. As ν_{it} is serially uncorrelated, the moment condition

$$E(x_{it-2}\Delta\nu_{it}) = 0 \quad (30)$$

validates instruments from time *t-2* and prior, which are then suitable for building a GMM estimator of equation (28) in first differences.

This estimator is problematic because highly persistent variables, like capital, produce weak instruments. Specifically, the first difference (Δx_{it}) will be poorly correlated with lagged levels (e.g., x_{it-2}), leading to significant bias in finite samples.

So, relying on Blundell and Bond (1998), if some other restrictions are set on the initial condition, another set of moments comes out:

$$E[x_{it-1}(\eta_i + \Delta v_{it})] = 0\,. \tag{31}$$

This result indicates that, to address endogeneity in the levels equation (28), lagged changes in the endogenous variables themselves as instrumental variables are used. The econometric method used combines the instruments suggested by moment conditions (30) and (31). To do this, it is necessary to put together the equations in differences (28) and the equations in levels (29). This combined approach gives reliable estimates of the coefficients, useful to find the original structural parameters in equation (26). For the purpose of examining the significance of biases related to fixed effects and endogeneity, the authors conduct estimations using random effects, within-groups, and GMM methods.

So, consider two complex issues: how data is grouped (aggregation) and whether training as a total amount (stock) or a rate (flow) is measured. Estimating at the industry level has advantages but also differs from estimating at the individual firm level. While equation (15) describes how training affects a single firm's productivity, theories of endogenous growth suggest that human capital benefits other firms too. For example, highly skilled workers might create new ideas that benefit the whole industry. This means equation (19) should include industry-wide training, and industry-level estimates should be larger than firm-level ones. Also, grouping by industry can reduce errors in the detailed firm-level data, improving accuracy.

Industry-level data offers benefits, but it can also create unpredictable aggregation biases. When higher-order moments remain constant across time, it is known that fixed effects will aid in mitigating some of these biases. Further problems arise, too, if coefficients differ haphazardly throughout enterprises. These changes are tested for in Dearden's empirical analysis. The data on the rate of training (flows) are known, but the model believes it has data on the total amount of training (stocks). So, it is possible to estimate the annual training rate objectively using the flow data. Alternatively, as described in Dearden 2005, the idea is to compute training stocks by applying a methodology similar to that of capital stocks, taking employee turnover into account as depreciation.

2.3 OTHER EMPIRICAL CONTRIBUTIONS: DEARDEN ET AL. RELATED STUDIES

The link between training practice and productivity has, for some years now, been a research topic in economic research as well as human capital theory. The landmark report by Dearden, Reed, and Van Reenen (2006) set an important precedent, showing that investment in employee training yields a greater return in productivity than wages, disproving the theory that pay increases will be able to

realize every benefit from improvements in personnel. Not only did this study put a figure on the impact of training — where 1% increase in training incidence led to a 0.6% increase in productivity and 0.3% increase in wages — but also stressed that endogeneity was important to ignore while studying the causal link between firm performance and training.

Drawing on this model, subsequent empirical studies have more recently sought to test if these findings apply more widely across national and industrial settings. Discussion below encompasses three key Italian, Belgian and US-based studies, in turn, that all take Dearden et al.'s model as a theory point of departure but translate it to the specific dynamics of its own national labor market.

Not only do these studies validate the core assumption that on-the-job training is more productive in increasing productivity than pay increases but also provide insight into how industry-specific circumstances, firm size, and human capital levels on hand affect the productivity of training programs. From a survey of these studies, we can gain a better understanding of how training methods interact with broader economic systems and have useful lessons for policymakers and firms.

The following is a comparative critical analysis of all the studies, their methods, findings, and points of congruence with Dearden et al.'s study. Comparison aims at presenting an overview of how training influences productivity on an international basis and setting up ground for evidence-informed methods for enhancing workforce development.

The Italian context.

Emilio Colombo and Luca Stanca's study, "*The Impact of Training on Productivity: Evidence from a Panel of Italian Firms*", which was published in the International Journal of Manpower (2014), uses a large panel dataset of Italian firms to examine how training affects labor productivity. Given that most of the earlier research concentrated mostly on the impacts of formal education, this study fills a vacuum in the economic literature about the accumulation of human capital through continuous learning.

The methodology

The authors assess the effect of training on labor productivity at the business level using an empirical method based on Dearden's econometric models. A Cobb-Douglas model is used to specify the production function, taking into account both trained and unskilled labor. To address any endogeneity issues of training, estimates are made using a variety of econometric techniques, such as the generalized method of moments (GMM), fixed effects (FE), and ordinary least squares (OLS). The dynamic GMM model was found to be the most suitable for capturing the causal relationship between

training and productivity, with special attention given to the potential for productivity shocks to impact businesses' training choices.

Dataset and data sources

Two primary data sources were combined to create the dataset used for the analysis:

- Excelsior: An annual survey by the Italian Ministry of Labor and Unioncamere that gathers comprehensive data on companies' training initiatives, such as the average length of training and employee participation in training programs (by job qualification: managers, white-collar workers, and blue-collar workers).
- The database AIDA[7] offers the yearly financial accounts of Italian companies with a turnover of more than 500,000 euros, considering factors like value added, capital, R&D expenses, and workforce size.

An unbalanced panel of 11,123 enterprises, including 33,815 observations for the years 2002–2005, makes up the final sample.

Main results

The empirical results indicate that training has a positive and significant impact on worker productivity at the firm level. In particular:

- On average, those with training are 10% more productive than those without training.
  - For so-called blue-collar workers, the impact of training is especially noticeable, as their output rises by 18%. On the other hand, the effect on managers and white-collar employees is statistically negligible.
  - The analysis emphasizes that underestimating the influence of training results from ignoring the length of training courses. The productivity effect is more than 20% when using a measure of effective training intensity, which is calculated by multiplying training intensity by the average number of training days per worker.
  - Manufacturing companies exhibit a less noticeable influence than other industries, but small businesses gain more from training than do medium and big businesses.

Comparison with the benchmark study

When comparing the findings of the Italian study with those of the British study "The Impact of Training on Productivity and Wages: Evidence from British Panel Data" (which served as baseline)

[7] Analisi Informatizzata Delle Aziende, Computerized Business Analysis

by Dearden, Reed, and Van Reenen (2006), it is possible to see that both studies support the idea that training increases firm productivity, but there are some notable differences:

- According to the British example, a 1% increase in training results in a 0.3% increase in earnings and a 0.6% gain in productivity.
- In Italy, instead, trained workers are 10% more productive than untrained workers, indicating a stronger impact that lower initial training levels may have in the Italian setting.
- The Italian study reveals significant heterogeneity: the impact is significant only for blue-collar workers, showing that less skilled individuals tend to profit more from continuous training, whereas the British study finds a pretty consistent effect across various occupational groups.

This discrepancy can be explained by taking into account that productive abilities are typically learned through formal educational pathways for managerial and white-collar jobs in Italy, which lessens the marginal benefit of firm-level training.

Note that Bratti, Conti, and Sulis did a study in Italy in 2018 and found out that when there's less training happening because companies use more temporary job contracts, it's a bummer for productivity. It turns out that whether bosses decide to train employees or not is affected by more than just how much extra money they think it'll make. It's also got to do with the whole job market scene out there.

The Belgian context.

Jozef Konings and Stijn Vanormelingen's paper, "*The Impact of Training on Productivity and Wages: Firm-Level Evidence*," which was published in the Review of Economics and Statistics in 2015, examines how workplace training affects wages and productivity at the firm level. In order to evaluate the implications of economic theories that explain firm-level training in the presence of imperfect labor markets, the primary goal is to determine whether and to what extent workplace training affects productivity differently than wages.

The methodology

The authors estimate production functions using firm-level panel data and a sophisticated econometric method based on the control function approach. Endogeneity, which occurs when businesses choose the degree of training in reaction to unobservable productivity shocks, is considered by this methodology.

The Cobb-Douglas specification governs the production function, which divides labor input between workers with and without training. In order to adjust for optimal input selections and produce consistent estimates for the influence of training on salaries and productivity, the Ackerberg, Caves, and Frazer (2006) method is used for the estimations. The production function is also estimated using a two-stage regression, and the pay equation is analyzed using total factor productivity as a control for unobservable workforce quality in the second stage.

Dataset and data sources

The two main sources of the dataset used are:

- The Belfirst Database is a for-profit resource made available by Bureau Van Dijck that includes financial accounts for Belgian businesses from 1997 to 2006. Value added, the number of employees (in full-time equivalent units), labor expenses, material costs, and capital stock are among the important elements included in the statistics.
- Formal training reports: in Belgium, businesses must provide an annual report detailing the number of staff members who have completed training, the hours spent on training, and the expenses incurred.

As a result, the authors were able to create accurate metrics for firm-level training, including the percentage of trained employees, the average number of training hours per employee, and training expenses.

Main results

The study results indicate that training has a statistically positive and significant effect on firm productivity, though with an even stronger impact than on wages:

- An increase of 10% in the proportion of trained workers is associated with a productivity increase of between 1.7% and 3.2%.
- In contrast, the average wage increase is lower, between 1.0% and 1.7%.
- Therefore, the training-induced productivity premium is almost double the wage premium.
- The effect of training is more prominent in non-manufacturing sectors than in manufacturing sectors.
- A trained worker has, on average, 32% higher productivity compared to an untrained worker, while their wage is merely 17% higher.

These results demonstrate the presence of wage compression in the Belgian labor market: firms will invest in general training despite the possibility that trained workers will change jobs, as productivity increases exceed wage increases.

Comparison with the benchmark study

The comparison with the British research by Dearden, Reed, and Van Reenen (2006) is helpful to highlight similarities and differences that are relevant:

- Both pieces of research confirm that training has a more significant impact on productivity than on wages. However, the British research estimates a 0.6% increase in productivity and a 0.3% increase in wages from a 1% increase in training, while the Belgian research identifies a stronger effect.
- Belgian firms seem to invest more in general training, pushing productivity ahead of wages at a faster rate, consistent with wage compression theory.
- While the British study is assessing training at sector level, the Belgian study adopts a microeconomic approach at firm level in order to analyse internal firm mechanisms in greater detail.

Overall, the Belgian study replicates the British study results, along with further empirical proof that the effect of training firms has on productivity is greater than on wages, in affirmation of the imperfect labor market theories of general training (Dearden et Al., 2006).

The United States context.

Facundo Sepulveda's article, "*Training and Productivity: Evidence for US Manufacturing Industries*", (2005) from Research School of Social Sciences at The Australian National University, discusses the relationship between training programs and productivity growth at the industry level in the US. The paper tries to identify the technology that transforms training into human capital and how it impacts aggregate productivity. Unlike earlier work, where productivity and wage impacts had often been confused with one another, Sepulveda demands clear-cut estimation of the productivity gains by means of production functions.

The methodology

The paper takes an econometric line of attack founded on the theory of production functions. The writer uses a Cobb-Douglas production function model, extended by the incorporation of human capital, comprising on-the-job training (OJT) and off-the-job training (OFFJT) as inputs. The model separates training as an investment in human capital rather than as a simple static stock. The core production function is specified as:

$$Y_t = A_t P(K_t, H_t \times L_t, M_t, E_t) \ , \tag{32}$$

where:

- Y_t: output at time
- A_t: stochastic productivity shock
- K_t: capital
- H_t: human capital
- L_t: labor input
- M_t: materials
- E_t: energy consumption

The accumulation of human capital is specified as a function of training, with current and past levels of training. Moreover, the paper enhances the model by checking the concavity of the training-productivity relation and the stability over time of the relations.

To address endogeneity — that the firms may change training levels due to unobserved productivity shocks — Sepulveda uses an instrumental variable (IV) approach with lags in the explanatory variables, as in Arellano and Bond (1991).

Data set and data sources

The sources include three major data sources:

- National Longitudinal Survey of Youth 1979 (NLSY79): Provides training data at the individual level, e.g., monthly training histories, which distinguish between on-the-job and off-the-job training. Formal training spells, training lengths, and completion rates are part of the data.
- The Board of Governors of the Federal Reserve System provides quarterly information on industrial production and electricity usage for United States manufacturing industries for the period 1988-1997. Bureau of Labor Statistics (BLS): Offers data on hours worked by industry.
- The resulting dataset is a quarterly panel of two-digit industrial sectors, from Q1 1988 through Q4 1997. Training variables were aggregated to the industry level by taking average monthly training incidences and spell completions.

Main results

The main findings of the study show subtle impacts of training on productivity growth:

- On-the-job training (OJT) makes a positive and sizable contribution to productivity growth. An increase in the incidence of training by 10% translates into an increase in productivity growth by 0.15 to 0.28 percentage points per quarter (depending on the method of estimation).

- Off-the-job training (OFFJT) has no measurable impact on productivity, implying that firm-specific skills learned through OJT are the only ones to directly result in increased output.
- The productivity effect of OJT is concave, i.e., returns to training fall as training intensity increases. The concavity parameter is estimated to be approximately -0.47.
- Human capital-intensive sectors have more returns to training. The interaction effect between OJT and the level of human capital (as measured by mean years of schooling in an industry) indicates that more educated sectors benefit disproportionately from training programs.
- There is no indication that the completion of a training spell, rather than just attending, yields extra productivity gains, contrary to some wage-based research which identifies pay increases for those employees who complete training programs. The longer-run productivity impact of training: Although training did not affect productivity at the start of the sample period (1988), by 1997, a 1% rise in training incidence raised productivity by 0.16% per quarter.

Comparison with the benchmark study

The findings of this study are also comparable to the findings of Dearden, Reed, and Van Reenen (2006) on the effects of training on productivity and wages in the UK. Some of the similarities are:

- Productivity impact sizes: both articles confirm that training boosts productivity above compensation. The UK study estimated a 0.6% productivity rise in response to a 1% rise in training incidence, while the US study estimates a 0.15 to 0.28 percentage point rise in quarterly productivity growth following a 10% rise in training incidence.
- Both exhibit the phenomenon of diminishing returns to training, and it explains that although initial investment in training yields enormous productivity growth, the advantages begin to decline with an escalation in training intensity.
- Sectoral differences: the US study emphasizes the link between training and human capital, with the finding that more educated sectors derive greater benefit from training. This was less elaborated in the British study, which was interested in average effects across sectors.
- Endogeneity adjustments: both studies employ variable instrumental techniques to deal with the endogeneity of training choices. However, the United States study uses more disaggregated quarterly data, allowing dynamic effects to be investigated in a more subtle manner.

In conclusion, Sepulveda's work presents firm evidence that on-the-job training has a considerable impact on productivity in U.S. manufacturing industries, with the highest impacts being in industries characterized by high stocks of human capital. The paper highlights the need to conceptualize training

as an investment in the accumulation of human capital, rather than as a fixed input, offering insight into the channels through which training affects aggregate productivity.
In contrast to the British comparison study, the United States study adds to understanding by studying variation in training effect over a broad selection of industries with varying levels of education and identifying time trends in training effectiveness. The data are crucial for policymakers who aim to create effective workforce training programs for specific industrial settings.

Comparative studies of training and productivity research conducted in different countries (specifically Italy, Belgium, United States and India[8]) demonstrates a positive and significant connection between on-the-job training and productivity improvement, corroborating the seminal findings formulated by Dearden, Reed, and Van Reenen (2006). The study by Dearden et al. made a significant theoretical contribution by providing empirical evidence that training raises productivity more than wages and, in doing so, also indicated the necessity of accounting for endogeneity in determining the causal effect of training on firm performance. Their study established the baseline that an increase of 1% in training frequency comes with a 0.6% increase in productivity and 0.3% in wages, thereby becoming a benchmark for subsequent studies in this area.

The findings of this research are utilized to restate the findings presented by Dearden et al., and to broaden their scope of relevance. The studies collectively highlight that training impacts are non-homogeneous but are powerfully influenced by industry-specific factors, labor market institutions, and the interaction between general and firm-specific skills. This cross-country evidence not only strengthens the link between training and productivity but also highlights the necessity to make training policies sector- and country-specific to fit the economic and educational contexts of different sectors and countries.
Prospectively, these results offer opportunities for further study. While the current literature successfully addresses problems of endogeneity and differential effects across industries, more analysis is possible on the long-term productivity trajectories of companies after training programs

[8] Singh and Mohanty's *Impact of Training Practices on Employee Productivity* (2010) study examines the connection between productivity and training in India's various industries. The study uses a comparative research design to gather secondary data from the Capitoline Plus Database and primary data via HR questionnaires. The results show that while training has little impact on high-risk or customer-driven industries like credit banks and luxury goods, it greatly increases productivity in stable industries like agriculture and automobiles. The result stands in contrast to a British study conducted by Dearden et al. (2006), which discovered that training consistently increased output across all industries. The Indian study uses comparative analysis and simulation approaches instead of the econometric approach used in the British study. In the end, it concludes that market variables unique to a certain industry have a significant impact on how effective training is.

as well as the interactive influence between organized training programs and vicarious learning through experience. Future comparative research could also examine to what extent larger economic events—such as recessions or technological progress—affect the link between training and productivity. Lastly, these studies emphasize that the development of human capital by way of targeted, evidence-driven training programs remains a valuable strategy for enhancing firm-level productivity and ensuring overall economic growth.

3. METHODOLOGY

The study's analytical methodology is covered in this section, along with how data from the INAPP survey were validated to determine how much employee productivity and overall business performance are improved by company training.

Instead of using longitudinal data, which follow changes over time like a series of snapshots, this study uses cross-sectional data. This method permits a thorough examination of the connection between production and training at a specific moment in time, even though it does not provide the detection of temporal changes.

The research was conducted using Stata, a well-known statistical program for econometric modelling. Using a multivariate linear regression technique, the study examines the connection between production and training expenditures. The approximated model looks like this:

$$productivity_i = \beta_0 + \beta_1(training_i) + \varepsilon_i \, . \tag{33}$$

Where:

- $productivity_i$: productivity of the i-th company (measured as output per hour);
- $training_i$: four metrics that quantify on-the-job training: total training spending, number of employees trained, percentage of employees trained, and training funding per employee;
- ε_i: error term.

This study uses a number of training-related variables to directly address the level of investment in training. The industry classification used the ATECO2007[9] coding method to further improve the estimation and ensure cross-sector comparability.

Robust regression (rreg) was chosen over ordinary least squares (OLS) estimate due to the dataset's characteristics. Because of its ideal characteristics under traditional assumptions, OLS is still a popular option, although it is extremely sensitive to anomalies in the data. There were outliers in the dataset, which are extreme but legitimate findings that show variations in business size, industry, and geographic distribution. Robust regression improves the dependability of estimates by giving extreme observations lower weights rather than eliminating these values, which can create selection bias.

The breach of OLS assumptions, specifically heteroscedasticity and non-normally distributed residuals, was another important justification for using robust regression. The Breusch-Pagan test and other diagnostic tests verified that the error variance varied among observations, which could result

[9] Obtainable at https://www.istat.it/wp-content/uploads/2022/03/volume_integrale_ATECO2007.pdf

in biased and ineffective standard errors in an OLS framework[10]. Heteroscedasticity-consistent standard errors do not completely resolve the impact of high-leverage points, but they may help to some extent. By iteratively modifying the model, robust regression guarantees more consistent and broadly applicable outcomes.

Iterative weighting was used in the estimation process; the Huber function first lessened the impact of big residuals, and then the bi-weight function improved the estimations. This methodological decision made sure that estimates of the coefficients captured the general trend in the data without being unduly impacted by extreme values.

Results from a comparison of robust regression and OLS showed how important this method is. OLS estimates occasionally produced surprising results, such as a negative correlation between productivity and training, because they were extremely sensitive to a small number of significant observations. The validity of statistical inferences was strengthened by robust regression, which yielded more consistent and stable coefficients.

The theoretical underpinnings of this research are derived from the widely reviewed publications of Dearden et al. (2006) and Becker (1964). These studies show that training and other expenditures in human capital benefit businesses by boosting production and employees by enhancing their abilities and earning potential. This viewpoint affected the interpretation of the empirical results and directed the choice of variables.

Lastly, the analysis finds connections between productivity and training but does not prove clear causal relationships because the study uses cross-sectional data. This restriction was carefully taken into account while interpreting the findings, highlighting the need for caution when drawing conclusions about potential policy consequences.

The complete Stata code used for the analysis is provided in the appendix under the title *Code.1*.

[10] See chapter XXX for results

4. STATA16 FOR THE RIL SURVEY

In this chapter, Stata16 is presented: the statistical analysis package that has been utilized for empirical analysis throughout this thesis. Stata is an extremely powerful, general, and simple software package that has been extensively used by economists, sociologists, political scientists, and a host of other researchers to data analyse, data manage, as well as graphically present. The following analysis is guided by the powerful econometric capabilities of Stata16, which is highly capable of managing complex data and yet possesses an amazing flexibility in conducting basic and advanced statistical procedures. That it is capable of both simple and advanced statistical modelling, in addition to being proficient in managing panel data, makes it extremely valuable to the empirical exercise carried out in this research.

4.1 AN OVERVIEW OF STATA

Stata is a statistical software package designed for statistical analysis, data visualization, data processing, and report generation. It has been a staple in social sciences for quantitative analysis tools since 1985. Its script-based interface assures that all components of analysis can be documented and repeated with precision, thereby enhancing the openness of the analytic process. This is crucial for academic research wherein replication of outcomes is a basic requirement. Apart from this, Stata covers a variety of subjects, from simple descriptives to sophisticated econometric techniques, making it possible for researchers to tailor their analyses to the needs of their studies.

This dissertation employs Stata16, which, compared to its predecessors, has more advanced features such as improved Bayesian econometrics, new panel-data features, and more developed integration with Python, which is helpful for more sophisticated data manipulation and visualization techniques. In addition, Stata16 added more tools for causal inference, which is especially useful in analysing the correlation between corporate training and productivity, as is done in this research. The software's ability to manage large datasets while offering reliable statistical tests made it a key component for the analysis performed.

4.2 THE LOGIC BEHIND STATA16

Choosing Stata16 was not random. It was informed by the nature of the dataset and the analytical needs of this thesis. For this analysis, INAPP RIL survey data was used, which contains information about training activities, workforce structure and productivity of Italian firms. Because of the panel nature of the dataset — the same firm is observed multiple times over a period — there was a requirement to use Stata's specific commands for panel data processing. These functions are essential

to longitudinal research for they allow the user to deal with unobserved heterogeneity and build dynamic panel models, which is necessary in establishing causal relationships in the data. Among the functions applied in this research were numerous of these programs at the same time, ranging from statistical analysis to the organization of the data. In particular, cleaning and structuring the RIL dataset required Stata16's data manipulation features merge, append, and reshape. An important precondition for effective analysis of the data has been that all variable values are correctly set and that their estimates are not missing, including methods that help mitigate biasing reasons.

Descriptive statistics such as `summarize`, `tabulate` and `list` have been used to help gain initial understanding of the dataset, as well as to shed light on the central variables such as training intensity, firm size and productivity levels. Moreover, the econometric functions of Stata16 made it possible to implement Ordinary Least Squares (OLS) models (regress). Despite the relative rigidity of Stata in data visualization in comparison to R or Python, Stata was known to be behind in this aspect. With the release of Stata16, there were some notable improvements. These examples serve to highlight that visualization is one of the critical steps in data analysis since it helps to expose structures and unusual patterns that would otherwise remain hidden.

4.3 ADVANTAGES AND LIMITATIONS OF STATA16

The primary benefits of utilizing Stata16 for this piece of research are as follows. First, Stata's user interface design is simplistic but quite potent. The program features tools with an easy-to-use layout and offers econometric functionalities with varying degrees of complexity depending on the user's level of programming experience. Second, Stata offers highly efficient functionality for managing and analysing panel data, even if it was not necessary for the firm within this research's longitudinal data analysis. Third, the software is effective in promoting reproducibility since with Stata's do-file system, processes of data analyses can be instructed and all data outputs stored in files for ease of reproduction – an important consideration in research.

It is also necessary to point out some limitations. Still, Stata has not maintained the level of graphical flexibility that most other software has, such as R or Python, which are easier to visualize. In addition, while the most basic of commands are usually easy, techniques such as GMM estimators or sophisticated data reshaping have their own learning curves. Finally, as with any proprietary nature, using the software is only available for license holders, which hinders a lot of researchers that work in limited resource scenarios.

To finish, Stata16 as a tool for this thesis proved to be invaluable in data management, statistical work, and general visualization of results. It was specifically designed to facilitate working with data and implementing econometric models for the study of the impact of corporate training on

productivity in Italy as stated in the INAPP RIL dataset. The blending of friendly interfaces, advanced statistical tools, and reproducibility ensured the empirical analysis, and the research was rigorous and transparent, adding credibility to the research findings.

5. DATA

The RIL survey, which INAPP performed, provided the dataset for this thesis. The RIL survey is a sample study that gathers comprehensive data on organizational strategy, corporate training, and work practices of Italian companies. Companies of all sizes and industries are included in the sample, which was chosen using a stratified sampling technique that guarantees national representativeness.

Corporate training is specifically covered in a section of the RIL questionnaire that collects information on the number of employees taking part in training activities, the kind of training (internal and external courses, on-the-job training, self-learning), and, for certain companies, the average length of training. The questionnaire collects information on corporate attributes including size, industry, region, and personnel makeup in addition to training data.

The RIL survey's reference population consists of operational businesses in all non-agricultural private sectors that are legally structured as partnerships or corporations and have no size constraints. As a result, single proprietorships, cooperatives, and other structures (consortia, organizations, etc.) are not included in the study. The reference population consists of 1,593,859 businesses and is taken from Istat's ASIA database (Statistical Archive of Active Enterprises), which was updated in 2013. Since the 2018 data utilized in this analysis is the most current accessible dataset, we infer that the same methodological approach applies even if the INAPP methodological comment relates to the 2015 survey.

An examination of the reference population was conducted prior to the sampling method in order to collect pertinent data for later stages of the study. A very successful sample and estimating technique was developed as a result of the preliminary investigation, leveraging the features that have the most impact on corporate behaviour during the design process. The Istat-provided ASIA 2013 database served as the survey's sampling frame.

The enterprise, which is a legal-economic entity that manufactures goods and services for the market, serves as the observation unit. This definition does not account for corporate ownership arrangements (group membership, mergers, spin-offs, etc.), stressing the legal-administrative aspect of the statistical unit. Through a special component of the questionnaire, the study gathered data on the ownership structure of businesses to identify factors influencing corporate strategy.

The survey questionnaire was first mailed out before the survey was carried out using the CATI (Computer-Assisted Telephone Interview) methodology. This approach guarantees a greater degree of process quality control by considering the resources at hand and the objective of carrying out an extensive nationwide survey. A letter signed by the Isfol President was sent to the chosen companies as part of the first communication. It contained information on the research project, survey goals, data gathering techniques, and data privacy protection, as well as the toll-free number and survey website.

The letter also referenced the survey's inclusion in the National Statistical Program and the obligation to respond. The questionnaire was enclosed to allow respondents, especially in larger and more structured firms, to gather the required information before the telephone interview, as specific sections of the questionnaire required input from various corporate departments (HR managers, general directors, sales managers, etc.).

Supplementary lists were employed to get the desired sample size and reduce the overall number of non-responses. These supplemental lists were used in accordance with recognized protocols designed to lessen respondent self-selection bias.

With variable probability selection of units and specified research domains, the sampling strategy uses a stratified sample design. Budgetary restrictions and the requirement for sufficiently precise estimates for certain subpopulations were balanced to establish the sample size, which was fixed at 30,000 units. Combining enterprise size (based on the average number of workers per year), the location of the company's legal headquarters, and the economic activity sector (based on an aggregate of the Ateco2007 categorization) allowed for stratification. After removing the population's empty layers, this produced 1,335 strata.

The purpose of sample allocation across strata was to provide a predefined degree of dependability for estimates pertaining to certain areas of interest, which were created by combining elementary strata. Alternative allocation techniques that consider minimal accuracy levels for specific areas of interest may be necessary since proportional allocation, although ensuring overall design efficiency, may not be sufficient to produce accurate estimates for tiny subpopulations. Using the Probability Proportional to Size (PPS) approach, which finds that the selection probability rises with firm size as shown by the average number of employees in 2013, the sample was selected from the ASIA database. When the firm size and the parameter of interest have a significant link, as has been shown in economic research, PPS sampling is more effective than constant probability sampling.

A coordinated sampling unit selection process based on the permanent random number approach was used to preserve the longitudinal sub-sample quota. By giving each population unit a random variable with a uniform distribution between 0 and 1, this technique ensures consistency between sampling occasions while maximizing sample overlap across many survey waves.

This study uses cross-sectional analysis with a single dataset gathered in 2018, in contrast to many economic analyses that use panel data to examine temporal fluctuations. This method enables the investigation of static correlations between corporate training and productivity, offering significant insights in line with the study goals, even though it does not capture changeable dynamics over time. There are benefits and drawbacks to this strategy. On the plus side, it provides comprehensive, organized data on a variety of Italian companies, giving a precise picture of the link between

productivity and training. Furthermore, the lack of longitudinal data prevents problems with non-response or sample attrition in later survey waves.
The inability to record the time-varying impacts of training on output is a major drawback, too. For instance, it is impossible to say if the effects of training show themselves gradually or only after a specific amount of time. This limits the ability to draw clear causal connections, thus results must be interpreted carefully.
Notwithstanding these drawbacks, cross-sectional analysis continues to provide insightful information regarding the relationships between corporate training and productivity, adding to the scholarly discussion and offering guidance for public and corporate human capital policy.

5.1 DATA ANALYSIS ON STATA

The INAPP RIL dataset captures information about specific companies' training, employment, and productivity. The first step involved importing the dataset into Stata using the `use` command, followed by some cleaning exercises.

```
use "/Users/melvinbellassai/Desktop/PUF RIL 05_18
stata/PUF_RIL_2018lavorato.dta"

estpost summarize
export excel using "Datafile", sheet("dirtyData") firstrow(variables)
```

With these commands Stata was asked to open the file containing data, then all the variables were summarized. The command `estpost summarize` returned descriptive statistics for the main variables of interest: mean, standard deviation, minimum and maximum values. It also revealed the data's overall structure. In the Appendix paragraph the output is reported, as *Table A.1*. Successively, data were exported to an Excel file, with the only objective to visualize properly some properties of the 227 variables (primarily negative values).
At this phase of the work, significant attention to detail was necessary since preparation mistakes could jeopardize the validity of all subsequent econometric analyses.

For the record, it is reported that a part of the data cleaning was carried out by the INAPP institute, as already stated. Specifically, consistency checks were already done. In the Excel file, only some unacceptable negative values came out. For this reason, these values were cancelled out, with the command:

```
drop if vB2<0 | vB4_5<0 | _v36<0 | _v44<0 | vC6_7<0 | vC10_3<0 | vF12<0 |
vL7<0 | vH12<0 | vH13<0
```

In addition, some completely useless[11] variables have been removed for the aim of this thesis.

```
drop vB3_VERIFICA vB4_VERIFICA
drop vB4BIS_1 vB4BIS_2 vB4BIS_2 vB4BIS_3 vB4BIS_4 vB4BIS_5
drop _v25 _v26 _v27 _v28 _v29 _v30 _v31 _v32 _v33 _v34 _v35 _v36
drop vC4_M_* vC4BIS
drop vC10_1 vC10_2 vC10_3
drop vC11_M_*
drop vC12 vC13_M_* vC14 vC14* vC15 vC15* vC16
drop vF7_M_* vF13 vF14
drop vL8_*
drop vH4_M_*
drop vH10_M_*
drop vH11
```

Where the * is used, it means that all the variables that begin with the characters before the asterisk and that end with anything else are taken by the command. Those variables were canceled out only because they are check variables, principally helpful for checking proposal, already done. Consequently, the variables in Tabel 1 below remained in the dataset.

VARIABLE	LABEL	MIS[12]	OBS[13]	MEAN	ST.DEV.[14]	MIN	MAX
CASENUM	Progressive ID	0	30003	15010,96	8667,33	1,00	30023,00
wcal	Weight	0	30003	52,63	229,81	0,00	5835,82
VA4	Legal form	0	30003	/	/	0,00	0,00
VA5	Year of incorporation (string)	0	30003	/	/	0,00	0,00
Ripartizione	Geographic breakdown	0	30003	/	/	0,00	0,00
ATECO2007	Ateco 2007 code	4	29999	/	/	0,00	0,00
vA9	Number of local units	0	30003	2,13	26,15	0,00	3773,00
vA11	Membership: YES/NO (string)	0	30003	/	/	0,00	0,00

[11] The variables in questions concern questions addressed to candidates to find out whether to proceed with subsequent questions or not; in essence, they are answers to “crossroad” questions.
[12] Number of missings
[13] Number of observations
[14] Standard deviation of the variable

_v1	Total employees	0	30003	57,96	327,86	0,00	36169,00
_v2	Tot. Empl. of which women	1	30002	20,88	162,28	0,00	17042,00
_v3	Tot. Empl. executives	4	29999	0,99	8,18	0,00	774,00
_v4	Tot. Empl. executives, of which women	4	29999	0,15	1,69	0,00	156,00
_v5	Tot. Empl. manager	4	29999	3,54	111,44	0,00	18377,00
_v6	Tot. Empl. managers, of which women	4	29999	1,06	41,84	0,00	6944,00
_v7	Tot. Empl. regular employees	0	30003	23,93	185,02	0,00	17018,00
_v8	Tot. Empl. regulars, of which women	0	30003	11,68	104,07	0,00	9983,00
_v9	Tot. Empl. workers	4	29999	29,43	142,26	0,00	9749,00
_v10	Tot. Empl. workers, of which women	3	30000	8,02	79,63	0,00	8291,00
vB2	N. of employees in 2016	4	29999	63,51	1519,53	0,00	256541,00
vB3_1	N. of Master's degrees	8665	21338	8,52	108,00	0,00	13273,00
vB3_2	N. of Bachelor's degrees	8665	21338	3,51	37,89	0,00	2731,00
vB3_3	N. of diploma's degrees	8665	21338	24,63	167,05	0,00	18956,00
vB3_4	N. of compulsory school	8665	21338	18,28	113,17	0,00	9592,00
vB3B_1	Distribution by Masters	24814	5189	8,94	16,38	0,00	100,00
vB3B_2	Distribution by Bachelor	24814	5189	5,85	14,18	0,00	100,00
vB3B_3	Distribution by diploma	24814	5189	49,26	29,06	0,00	100,00
vB3B_4	Distribution by compulsory school	24814	5189	35,95	31,51	0,00	100,00
vB4_1	Distribution <25 y.o.	7208	22795	2,81	18,81	0,00	900,00
vB4_2	Distribution 25-34 y.o.	7209	22794	11,22	54,54	0,00	2834,00
vB4_3	Distribution 35-49 y.o.	7208	22795	27,29	168,48	0,00	17542,00
vB4_4	Distribution 50-59 y.o.	7208	22795	15,58	120,13	0,00	14208,00
vB4_5	Distribution >60 y.o.	7208	22795	2,78	21,12	0,00	1634,00

_v11	Tot. permanent empl.	3474	26529	56,72	324,37	0,00	35301,00
_v12	Tot. permanent empl., of which women	3473	26530	20,16	157,44	0,00	16565,00
_v13	Tot. Fixed-term empl.	3473	26530	6,58	36,81	0,00	1616,00
_v14	Tot. Fixed-term empl., of which women	3473	26530	2,56	20,25	0,00	1259,00
_v15	Tot. apprenticeship	3473	26530	1,39	14,12	0,00	927,00
_v16	Tot. apprenticeship, of which women	3473	26530	0,52	7,51	0,00	599,00
_v17	Tot. Job-on-call	3473	26530	0,82	41,35	0,00	5725,00
_v18	Tot. Job-on-call, of which women	3473	26530	0,41	25,70	0,00	3911,00
_v19	Tot. part-time empl.	3471	26532	11,03	110,97	0,00	8890,00
_v20	Tot. part-time empl., of which women	3471	26532	7,92	91,25	0,00	8125,00
_v21	Tot. permanent part-time empl.	3471	26532	9,24	97,75	0,00	7933,00
_v22	Tot. permanent part-time empl., of which women	3471	26532	6,83	83,07	0,00	7329,00
_v23	Tot. fixed-term part-time empl.	3471	26532	1,80	20,33	0,00	1242,00
_v24	Tot. fixed-term part-time empl., of which women	3471	26532	1,08	13,13	0,00	796,00
vB10	Training hours organized? YES/NO	0	30003	/	/	0,00	0,00
vB11	Number of trained employees	14178	15825	60,89	389,48	0,00	37843,00
_v37	Permanents hired before 01/2017	18877	11126	52,51	406,05	0,00	37006,00
_v38	Permanents hired before	18877	11126	17,16	183,03	0,00	17152,00

	01/2017, of which women						
_v39	Permanents hired after 01/2017	18877	11126	8,98	301,88	0,00	30123,00
_v40	Permanents hired after 01/2017, of which women	18877	11126	2,16	33,91	0,00	1946,00
_v41	Fixed-term hired before 01/2017	18877	11126	2,19	15,85	0,00	522,00
_v42	Fixed-term hired before 01/2017, of which women	18877	11126	0,75	8,16	0,00	340,00
_v43	Fixed-term hired after 01/2017	18877	11126	3,69	27,01	0,00	1250,00
_v44	Fixed-term hired after 01/2017, of which women	18877	11126	1,48	17,51	0,00	1125,00
vB12_M_1	Type of training: support	14169	15834	/	/	0,00	0,00
vB12_M_2	Type of training: compulsory	14169	15834	/	/	0,00	0,00
vB12_M_3	Type of training: technical	14169	15834	/	/	0,00	0,00
vB12_M_4	Type of training: computer	14169	15834	/	/	0,00	0,00
vB12_M_5	Type of training: other	14174	15829	/	/	0,00	0,00
vB13	Costs incurred in full or with some funds?	14169	15834	/	/	0,00	0,00
vB14_M_1	Type of fund: social	24722	5281	0,70	0,46	0,00	1,00
vB14_M_2	Type of fund: state	24722	5281	0,27	0,44	0,00	1,00
vB14_M_3	Type of fund: other	24722	5281	0,11	0,31	0,00	1,00
vB15_2	Expenses for training	18589	11414	77347,79	2241632,08	0,00	200000000,00
vD1	Fixed-time use reason (string)	17421	12582	/	/	0,00	0,00
vD2	Part-time use reason (string)	11739	18264	/	/	0,00	0,00
vD3	Apprenticeship use reason (string)	24002	6001	/	/	0,00	0,00

vD4	Collaboration use reason (string)	20193	9810	/	/	0,00	0,00
vD5	Temporary job use reason (string)	26565	3438	/	/	0,00	0,00
vC1	Hiring made in 2017? YES/NO	0	30003	/	/	0,00	0,00
vC2	N. of hired in 2017	14442	15561	17,76	104,66	0,00	9234,00
vC3	Hires in 2017 using public funds	14377	15626	/	/	0,00	0,00
vC5	Termination of work in 2017? YES/NO	0	30003	/	/	0,00	0,00
vC6_1	Tot. terminations	14718	15285	16,83	106,57	0,00	9246,00
vC6_2	Terminations type: layoffs	14718	15285	1,71	10,94	0,00	378,00
vC6_3	Terminations type: retirements	14718	15285	0,78	3,68	0,00	198,00
vC6_4	Terminations type: pre-retirements	14718	15285	0,33	26,85	0,00	3304,00
vC6_5	Termination type: end of term contracts	14718	15285	8,65	86,85	0,00	8572,00
vC6_6	Terminations type: resignations	14718	15285	4,26	17,75	0,00	736,00
vC6_7	Terminations type: other	14718	15285	1,10	14,07	0,00	713,00
vC7	Looking for staff? YES/NO	0	30003	/	/	0,00	0,00
vC7BIS	N. of people looked for	25056	4947	5,45	17,97	0,00	500,00
vC8_M_1	Profile type looked for: manager	25037	4966	/	/	0,00	0,00
vC8_M_2	Profile type looked for: high specialist	25037	4966	/	/	0,00	0,00
vC8_M_3	Profile type looked for: technical	25037	4966	/	/	0,00	0,00
vC8_M_4	Profile type looked for: office staff	25037	4966	/	/	0,00	0,00
vC8_M_5	Profile type looked for: services	25037	4966	/	/	0,00	0,00
vC8_M_6	Profile type looked for: skilled workers	25037	4966	/	/	0,00	0,00

vC8_M_7	Profile type looked for: non-skilled workers	25037	4966	/	/	0,00	0,00
vC8_M_8	Profile type looked for: other	25037	4966	/	/	0,00	0,00
vC9	Labor intermediaries used? YES/NO	0	30003	/	/	0,00	0,00
vF1	Membership in a trade association? (string)	0	30003	/	/	0,00	0,00
vF1BIS	N. of memberships	28282	1721	2,19	0,67	0,00	11,00
vF2	Collective agreements application	3471	26532	/	/	0,00	0,00
vF3	Collective agreement type	8124	21879	1,07	0,26	1,00	2,00
vF3_2	Collective agreement type (text)	0	30003	/	/	0,00	0,00
vF3B	Description of the agreement	0	30003	/	/	0,00	0,00
vF4	Second-level bargaining done? YES/NO	3468	26535	/	/	0,00	0,00
vF5	Type of second-level barg.	26747	3256	/	/	0,00	0,00
vF6_M_1	Type of SLB[15]: result awards	26747	3256	/	/	0,00	0,00
vF6_M_2	Type of SLB: working hours	26747	3256	/	/	0,00	0,00
vF6_M_3	Type of SLB: training	26747	3256	/	/	0,00	0,00
vF6_M_4	Type of SLB: equal opport.	26747	3256	/	/	0,00	0,00
vF6_M_5	Type of SLB: health care	26747	3256	/	/	0,00	0,00
vF6_M_6	Type of SLB: labor market	26747	3256	/	/	0,00	0,00
vF6_M_7	Type of SLB: environment	26747	3256	/	/	0,00	0,00
vF6_M_8	Type of SLB: welfare	26747	3256	/	/	0,00	0,00
vF6_M_9	Type of SLB: contractual minimums	26747	3256	/	/	0,00	0,00
vF6_M_10	Type of SLB: participation in company decisions	26747	3256	/	/	0,00	0,00

[15] SLB: second-level bargaining

vF6_M_11	Type of SLB: other	26747	3256	/	/	0,00	0,00
vF10	Union representation presence? YES/NO	3471	26532	/	/	0,00	0,00
vF11	N. of workers in a union	24560	5443	68,21	449,37	0,00	28219,00
vF12	Strike hours	4821	25182	1542,02	233384,29	0,00	37034000,00
vL1	Product/service innovations applied 2015/17? YES/NO	0	30003	/	/	0,00	0,00
vL2	Production process innovations applied 2015/17? YES/NO	0	30003	/	/	0,00	0,00
vL3	Patents purchased in 2015/17? YES/NO	0	30003	/	/	0,00	0,00
vL4_1	Investments in: IoT (YES/NO/FUTURE)[16]	0	30003	/	/	0,00	0,00
vL4_2	Investments in: robotics	0	30003	/	/	0,00	0,00
vL4_3	Investments in: Big Data Analytics	0	30003	/	/	0,00	0,00
vL4_4	Investments in: VR/AR	0	30003	/	/	0,00	0,00
vL4_5	Investments in: Cybersecurity	0	30003	/	/	0,00	0,00
vL4_6	Investments in: updating devices	0	30003	/	/	0,00	0,00
vL4_7	Investments in: other	0	30003	/	/	0,00	0,00
vL6	Export? YES/NO	0	30003	/	/	0,00	0,00
vL7	% of exports	22464	7539	38,05	138,89	0,00	11412,00
vL9	All or part of the production exported?	0	30003	/	/	0,00	0,00
vL10	PA supplier? YES/NO	354	29649	/	/	0,00	0,00
vL11	Revenues from PA in %	22188	7815	22,95	30,82	0,00	100,00
vH1	Investments done? YES/NO	443	29560	/	/	0,00	0,00

[16] Answers applicable to variables from vL4_1 to vL4_7

vH2	Total investments in €	19027	10976	4428315,07	1,18E+08	0,00	9,30E+09
vH3	Use of incentives for investments?	17430	12573	/	/	0,00	0,00
vH5_M_1	Type of inv.: marketing	17430	12573	/	/	0,00	0,00
vH5_M_2	Type of inv.: R&D	17430	12573	/	/	0,00	0,00
vH5_M_3	Type of inv.: lands	17430	12573	/	/	0,00	0,00
vH5_M_4	Type of inv.: general equipment	17430	12573	/	/	0,00	0,00
vH5_M_5	Type of inv.: computer equipment	17430	12573	/	/	0,00	0,00
vH5_M_6	Type of inv.: other	17430	12573	/	/	0,00	0,00
vH6	Bank credit request for liquidity? YES/NO	0	30003	/	/	0,00	0,00
vH7	Bank credit outcome? String[17]	22749	7254	/	/	0,00	0,00
vH8	Bank credit request for investments? YES/NO	0	30003	/	/	0,00	0,00
vH9	Bank credit outcome? String[18]	24643	5360	/	/	0,00	0,00
vH12	Total hours worked	10897	19106	227336,96	5973760,10	0,00	705973000,00
vH13	Total revenues in 2017	4059	25944	4,19E+10	5,20E+12	17,00	7,83E+14

Table 1: varibles description

Selection and presentation of key variables.

A thorough explanation of the important variables that will be used for this thesis comes next, after the data cleaning step. As *Table 1* illustrates, the RIL dataset includes many variables, but not all of them are pertinent to the particular research questions this study aims to answer. Thus, only the factors that are directly associated with employee productivity, corporate training, and business characteristics have been chosen for more examination.

The chosen variables will be shown separately, using the Stata command

[17] Answers: totally granted, partially granted, not granted
[18] Answers: totally granted, partially granted, not granted

```
summarize variable, detail
```

which offers a full summary of each variable's distribution, including not only the number of observations, mean, standard deviation, and range (minimum and maximum values), but also percentiles, variance, skewness and kurtosis statistics. This step is essential since it guarantees the precision and dependability of ensuing econometric studies and permits a comprehensive grasp of the data's structure.

These instructions' output will be closely analysed and debated, with special attention paid to any noteworthy trends or anomalies that could affect how the results are interpreted. Missing data, outliers, and unusual distributions will receive special attention since they can have a big impact on how resilient the models are.

The variables will be presented in a logical order, beginning with those that measure the intensity of corporate training, then productivity, and lastly control factors like sector and type of OJT. To give a thorough overview of the dataset, a brief description of each variable will be included, along with commentary on its descriptive statistics.

Total number of employees

So, from now on, the command used is like this form:

```
summarize _v1, detail
```

In this specific case the variable analyzed is `_v1`. The output is reported below in *Table 2*.

LAVORATORI DIPENDENTI_Totale

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	30,006
25%	3	0	Sum of Wgt.	30,006
50%	11		Mean	57.95744
		Largest	Std. Dev.	327.8701
75%	38	10178		
90%	111	10852	Variance	107498.8
95%	209	11850	Skewness	54.04055
99%	755	36169	Kurtosis	5133.048

Table 2: variable _v1 description

The variable `_v1` refers to the number of employees in each organization. Descriptive statistics indicate numerous key elements of its distribution. The average number of workers is 57.96, with a maximum of 36,169, showing a highly skewed distribution. This shows that, while most businesses

are tiny, a few major ones greatly elevate the average. *Figure 1* shows that in a better way[19]. The median of 11 confirms this pattern, indicating that half of the organizations in the sample had 11 or fewer employees. The difference between the mean and median highlights the presence of outliers, with a small number of companies employing a disproportionately big staff. The results indicate that the distribution is highly right skewed, with a skewness score of 54.04.

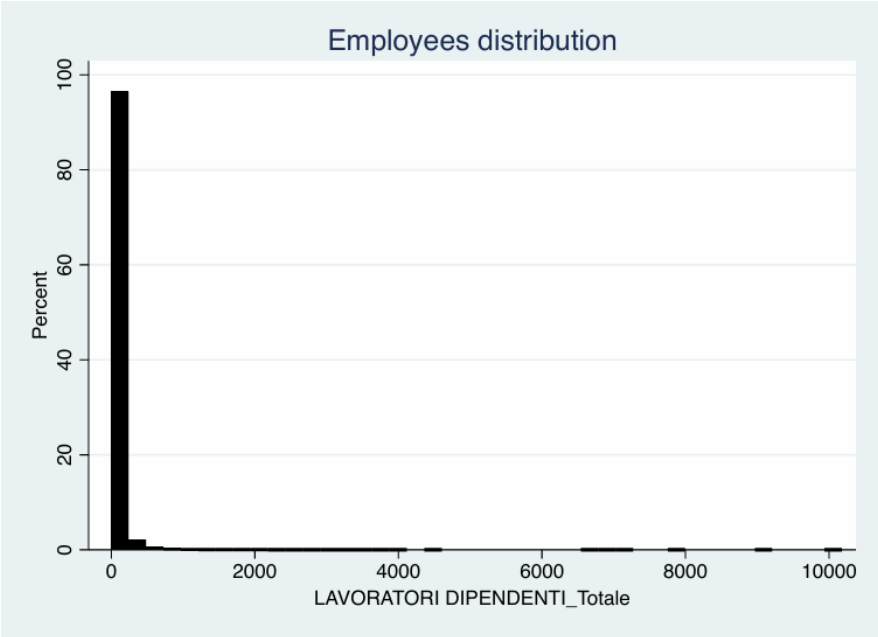

Figure 1: employees distribution

This severe imbalance suggests that the bulk of businesses employ considerably fewer people than a small number of really large corporations. The kurtosis value of 5133.05 supports this finding by suggesting a distribution with a strong peak and extended tails, which is characteristic of datasets containing extreme values. Notably, although the 99th percentile has 755 people, the largest business has a headcount of 36,169, highlighting the presence of significant outliers. The standard deviation, 327.87, is much higher than the mean, indicating strong variability in firm size. Such broad dispersion indicates considerable disparities in the sample, with tiny enterprises coexisting with giant multinationals. This fluctuation has significant implications for the ensuing econometric study.

Furthermore, the existence of severe outliers may skew the findings of regression models. As a result, special care will be taken to assess the influence of these huge enterprises and ensure that they do not have an undue impact on the projections.

Overall, the descriptive statistics of `_v1` highlight a dataset characterized by a majority of small companies and a minority of significantly larger ones, a pattern that reflects the well-known structure of the Italian corporate landscape (Costa, S., De Santis, S., & Monducci, R., 2022). According to this

[19] The graph is obtained with:

```
histogram _v1, title("Employees distribution") color(black) percent
```

study, Italy is dominated by small and medium-sized enterprises (SMEs), with a relatively small proportion of large firms, which aligns with the findings of the present dataset.

Total sales for 2017

Asking Stata to analyze the variable `vH13`, always using the command `summarize` with the `detail` addiction, the output is:

Facendo riferimento al bilancio dell'impresa del 2017, qual è l'ammontare d

	Percentiles	Smallest		
1%	10000	17		
5%	60000	30		
10%	117600	30	Obs	25,947
25%	356569	50	Sum of Wgt.	25,947
50%	1600000		Mean	4.19e+10
		Largest	Std. Dev.	5.20e+12
75%	7536595	4.89e+11		
90%	3.10e+07	5.63e+12	Variance	2.70e+25
95%	7.54e+07	2.97e+14	Skewness	138.7864
99%	4.12e+08	7.83e+14	Kurtosis	20224.51

Table 3: variable vH13 description

The variable `vH13` represents the total revenue of Italian companies for the year 2017. The descriptive statistics reveal a distribution characterized by a significant skewness and the presence of extreme outliers. Specifically, the 1st percentile of revenue is 10,000 euros, while the 99th percentile reaches a substantial 412 million euros, with the largest value at 783 billion euros. The 25th percentile is 356,569 euros, the 50th percentile (median) stands at 1.6 million euros, and the 75th percentile is at 7.5 million euros. This indicates that half of the companies in the dataset have revenues below 1.6 million euros, and a quarter of them have revenues below 356,569 euros.

Since according to the data provided by Mediobanca and reported by Panorama[20], the company with the highest revenue in Italy in 2017 was Enel Italia, with 73 billion euros, a value of 783 trillion euros is not acceptable. Maybe, the recorder recorded wrongly the number of zeros, or the data enter did so. Therefore, it was decided to make the values of vH13 greater than 100 billion missing using the following command:

```
replace vH13 = . if vH13>1e11
```

At this point, it is possible to redo the detailed analysis of the variable, which should now be more reliable. The new output is reported in *Table 4*.

[20] La classifica delle imprese italiane per fatturato - Panorama

Facendo riferimento al bilancio dell'impresa del 2017, qual è l'ammontare d

	Percentiles	Smallest		
1%	10000	17		
5%	60000	30		
10%	117600	30	Obs	25,941
25%	356564	50	Sum of Wgt.	25,941
50%	1600000		Mean	3.65e+07
		Largest	Std. Dev.	6.94e+08
75%	7527000	2.51e+10		
90%	3.09e+07	2.63e+10	Variance	4.82e+17
95%	7.50e+07	4.52e+10	Skewness	72.36175
99%	4.03e+08	7.70e+10	Kurtosis	6806.979

Table 4: variable vH13 (cleaned) description

Having replaced the unacceptable data, nothing changes in percentiles reported before. The only change is on the maximum: six values were replaced with the missing value. The new maximum is 77 billion, that is more acceptable.

At 36.5 million euros, the mean revenue is much greater, indicating that a small number of extremely large companies are having a substantial impact on the average. The fact that there are outliers in the data, with a few very large enterprises inflating the general mean, is shown by this difference between the mean and the median.

Significant variations in income throughout the sample are shown by the variance of 4.82 × 10^17 and the very high standard deviation of 694 million euros. A significant right skew in the data is confirmed by the skewness value of 72.36, which indicates that although most businesses have relatively low revenues, a few extremely large businesses drive the distribution toward higher revenue values.

This finding is further supported by the kurtosis value of 6,806.98, which indicates that the distribution has a strong peak and extended tails to the right, which is common in datasets containing severe outliers.

The sample's heterogeneity (a considerable percentage of businesses are tiny, while a small number of major organizations dominate in terms of overall revenue) is highlighted by the high skewness and variability in revenue, as visible in *Figure 2*[21].

[21] The graph is obtained with:
`histogram vH13, title("Total sales distribution (2017)") color(black) percent`

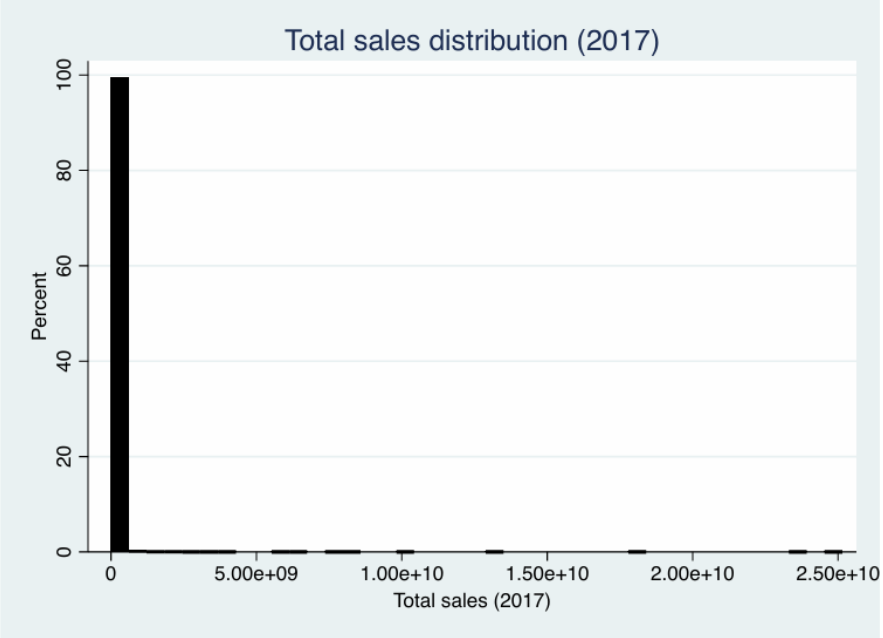

Figure 2: total sales distribution (2017)

These results are in line with the overall composition of the Italian business environment, which is dominated by small and medium-sized businesses (SMEs), with a few giant corporations holding an excessively large amount of the total revenue.

Number of graduates in the company

The analysis will now shift to the companies' labour makeup, looking at four different variables: `vB3_1`, `vB3_2`, `vB3_3`, and `vB3_4`, respectively the number of employees with a master's degree, bachelor's degree, high school diploma, and those who only complete the mandatory education. Examining the workforce's educational background will reveal important information about the organizations' human capital structure and the possible correlation between educational achievement and business performance. Since the question about the distribution of employees through their education is asked only if the number of employees (i.e. `_v1`) is higher than 0, the values entered for companies without employees is the missing one. To not lose information, commands below are asked to Stata.

```
replace vB3_1=0 if _v1==0
replace vB3_2=0 if _v1==0
replace vB3_3=0 if _v1==0
replace vB3_4=0 if _v1==0
```

Before presenting the descriptions of these variables it appears necessary to check and drop the variable corresponding to the firm if the sum of the four in exam here is higher than the total number

of employees (i.e. `_v1`), or lower than the it. It is possible also to say that the sum equals `_v1`. With this aim, it is demanded to Stata to do

```
generate sum_vB3= vB3_1+vB3_2+vB3_3+vB3_4
drop if sum_vB3>_v1 | sum_vB3<_v1
```

In that way, 5,194 observations are deleted. Most of them are deleted because of Stata considers the missing value of `sum_vB3` higher than `_v1`. There were 5,194 missing values generated for `sum_vB3` because of missing values recorded in the four variables that compose the sum. It is not possible to trace why these missing values were recorded, therefore it is necessary to delete them.

At this point, we can analyse the four variables singularly.

The number of workers in each company who hold a master's degree is represented by the variable `vB3_1`. As *Table 5* shows, there are 24,812 observations in the dataset. There is a notable concentration of businesses with few or no master's degree holders in the distribution; the first, fifth, tenth, twenty-five, and even fiftieth percentiles are all at zero, meaning that at least half of the businesses do not employ anyone with a master's degree. Just 25% of companies have two or more master's grads, according to the 75th percentile. With a maximum value of 13,273, the upper end of the distribution shows more fluctuation, with the 90th percentile at 9, the 95th at 24, and the 99th at 117 master's degree holders. The standard deviation of 100.21 and variance of 10,041.79 indicate that a small number of organizations with extraordinarily high counts have a significant influence on the mean number of master's degrees per company, which is 7.33.

A strongly right-skewed distribution is shown by the skewness value of 99.08, which confirms that a small percentage of enterprises employ a disproportionately large number of master's degree holders, while the majority have very few master's degree holders.

Distribuzione lavoratori dipendenti per titolo studio_Laurea/Laurea Magistrale o

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	24,812
25%	0	0	Sum of Wgt.	24,812
50%	0		Mean	7.329316
		Largest	Std. Dev.	100.2087
75%	2	2324		
90%	9	3445	Variance	10041.79
95%	24	3697	Skewness	99.08101
99%	117	13273	Kurtosis	12537.5

Table 5: variable vB3_1 description

The distribution's extreme outliers and heavy tails are further highlighted by the kurtosis value of 12,537.5. Any econometric research incorporating this variable may necessitate a customized strategy, such as a zero-inflated model or a transformation of the data to better handle the large range of values, due to the extreme skewness and concentration of zeros. Even if most organizations only hire a small number of highly educated people, it is nevertheless helpful to understand the distribution of master's degree holders in order to evaluate the human capital structure within firms.

The number of workers with a bachelor's degree in each company is represented by the variable `vB3_2`, described in *Table 6*. There are 24,812 observations in the dataset. The data reveal a high concentration of companies with few or no bachelor's degree holders, like the distribution of master's graduates: the first, fifth, tenth, twenty-five, and fiftieth percentiles are all at zero, meaning that at least half of the companies do not employ any people with a bachelor's degree, similarly to `vB3_1`.
With a maximum value of 2,731, the 75th percentile stays at 0, the 90th percentile increases to 3, the 95th to 8, and the 99th to 50 bachelor's grads.
Although a small number of companies employ significantly more degree holders, the mean number of bachelor's graduates per company is 3.02. With a variance of 1,236.35 and a standard deviation of 35.16, the sample exhibits significant dispersion.
A substantially right-skewed distribution is revealed by the skewness value of 41.52, indicating that a small group of enterprises employ a significantly higher number of bachelor's degree holders than the majority, which employ few or none at all. The distribution's lengthy tails and strong outliers are further highlighted by the kurtosis of 2,369.18.
As with master's grads, the extreme skewness and number of zeros indicate that a zero-inflated technique or data transformation could be helpful for handling the large range of values in any econometric modelling incorporating this variable.

Distribuzione lavoratori dipendenti per titolo studio_Diploma uni./Laurea trienn

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	24,812
25%	0	0	Sum of Wgt.	24,812
50%	0		Mean	3.016202
		Largest	Std. Dev.	35.16179
75%	0	1489		
90%	3	1600	Variance	1236.352
95%	8	1945	Skewness	41.5204
99%	50	2731	Kurtosis	2369.182

Table 6: Variable vB3_2 description

Even though many businesses do not hire anyone with a bachelor's degree, knowing how these graduates are distributed is essential to determining a company's human capital profile and how it might affect its performance.

The number of workers having a high school degree in each company is represented by the variable `vB3_3`, in *Table 7*. There are 24,812 observations in the dataset. The distribution of high school graduates among firms is somewhat more widespread than that of master's and bachelor's degree holders, according to the data. The median (50th percentile) is at 3, indicating that at least half of the companies have a small but noticeable number of employees with a high school diploma, while the first, fifth, and tenth percentiles stay at 0, suggesting that some firms employ no high school graduates.

Distribuzione lavoratori dipendenti per titolo studio_Diploma di scuola media su

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	24,812
25%	1	0	Sum of Wgt.	24,812
50%	3		Mean	21.18636
		Largest	Std. Dev.	155.1672
75%	12	3975		
90%	38	4542	Variance	24076.87
95%	78	5540	Skewness	79.13295
99%	291	18956	Kurtosis	9079.787

Table 7: variable vB3_3 description

The upper part of the range has a lot more differences: the 75th spot gets up to 12, but the 90th is like way higher at 38, and the 95th goes even crazier to 78. Then, the 99th percentile is at 291, and the absolute highest is a massive 18,956.

For the average number of high school grads per company, it's 21.19, which is a big deal when you compare it to the master's peeps, who are at 7.33, and the bachelor's folks at 3.02. This basically means that high school grads are like most of the workers out there. There's also a lot of spread here, with a standard deviation of 155.17 and a variance of 24,076.87, which tells that companies really vary when it comes to the number of employees with different education levels.

The skewness score of 79.13 indicates a strongly right-skewed distribution, even if slightly less extreme than for master's and bachelor's grads, while the kurtosis of 9,079.79 indicates the existence of extreme outliers and lengthy tails.

Compared to the previous variables, the distribution of high school graduates indicates a more common, but unequal, presence of mid-level education in the workforce. While master's and

bachelor's graduates appear to be concentrated in a small number of organizations, high school graduates are more evenly distributed, with some noticeable outliers.
This tendency is consistent with broader labor market trends, in which intermediate education levels are often the foundation of many enterprises' workforces. As Ferri, V., Ricci, A., & Sacchi, S. (2018) point out, human capital distribution within enterprises frequently parallels national education trends, with advanced degrees concentrated in specialized areas and high school graduates spread over a wider range of industries.

The variable `vB3_4` reflects the number of employees who have only finished obligatory education, and the dataset has a total of 24,812 observations. The results, presented in *Table 8*, show a high concentration of employees with low or zero values, similar to the distribution of other variables in which the majority of employees lack higher educational attainment. The first, fifth, tenth, twenty-fifth, and fifty percentiles are all zero, indicating that at least half of the companies has no employees or employees that have not finished anything beyond compulsory education. The 75th percentile is 8, the 90th is 30, the 95th is 60, and the 99th is 220, indicating that, while the vast majority of employees have low values, some observations have significantly higher values.
The mean of 15.72 is relatively low, but the high standard deviation (105.15) and variance (11,055.79) indicate significant dispersion, with some values deviating considerably from the mean.

Distribuzione lavoratori dipendenti per titolo studio_Scuola dell'obbligo

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	24,812
25%	0	0	Sum of Wgt.	24,812
50%	1		Mean	15.72231
		Largest	Std. Dev.	105.1465
75%	8	4071		
90%	30	4108	Variance	11055.79
95%	60	6357	Skewness	48.2555
99%	220	9592	Kurtosis	3558.919

Table 8: Variable vB3_4 description

The skewness value of 48.26 confirms a strongly right-skewed distribution, indicating that a tiny percentage of employees have significantly higher values than the majority. The exceptionally high kurtosis (3,558.92) indicates the presence of heavy tails, which means that there are many extreme values that have a large impact on the entire distribution.
In conclusion, the distribution of this variable is characterized by a great majority of low values, with a tiny fraction of employees having significantly higher values, resulting in a strongly skewed

distribution. Despite the prevalence of individuals with lower educational levels, understanding how they are dispersed is critical since it may have ramifications for a company's human capital profile and possible impact on performance.

Number of employees trained on the job

Before studying `vB11`, it's crucial to know how many observations were taken, because not every organization offers training. The command

```
count if missing(vB11)
```

instructs Stata to count and display the number of missing values for the variable in question. There are 12,525 missing values. Because we do not want to lose this information, we must determine where the missing data are coming from. A check of the questionnaire reveals that the variable `vB10` contains the replies to the question: "Were training initiatives organized for the company's employees in 2017?". The appendix section contains, as *Code.2*, the code used to retrieve the description of the vB10 variable: the answer "No" corresponds to the value "2". As a result, we replace the missing values in `vB11` with zero for companies who did not give training:

```
replace vB11 = 0 if vB10 == 2
```

Because the question about the number of employees that participated in training was only asked of enterprises with a positive number of employees, we update `vB11` by replacing the missing values if the company in issue has no employees.

```
replace vB11 = 0 if _v1 == 0
```

Stata reports modifying 80 observations (see *Code.3* in the appendix), which is a nearly acceptable number.

Another critical step is to identify and remove instances in when the number of trained employees exceeds the total number of employees. As previously said, the following command is utilized:

```
count if vB11 > _v1
```

The `count` command is not necessarily necessary, but it does help to ensure that the cleaning process is carried out appropriately. The count reports 370 observations, which is still a reasonable number:

just a little bit over 1% of the total observations. If this number had been substantially higher, it would have been required to investigate where the error originated and reassess the entire operation. The `drop` command deleted the 370 observations (see *Code.3* in the appendix):

```
drop if vB11 > _v1
```

At this point, the description of the variable `vB11` can be done, and the results are presented in *Table 9*.

Quanti dipendenti dell'impresa hanno partecipato complessivamente a tali inizi

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	24,442
25%	0	0	Sum of Wgt.	24,442
50%	0		Mean	22.60813
		Largest	Std. Dev.	138.0862
75%	10	5027		
90%	40	5107	Variance	19067.79
95%	88	5399	Skewness	31.2014
99%	358	10178	Kurtosis	1624.035

Table 9: variable vB11 description

The variable vB11, as stated, reflects the number of employees who attended training, with a total of 24,442 observations. The code presented before replaced missing numbers with zero for companies that did not provide training in 2017. This adjustment results in a variable that is significantly concentrated around zero, as many organizations did not provide any training, and those that did frequently have a small number of trained personnel. *Figure 3* displays it better[22].

Looking at the percentiles, the 1st, 5th, 10th, and 25th percentiles are all zero. This suggests that at least a quarter of the organizations did not have any employees enrolled in training, which supports our initial estimate that many companies did not provide training.

The median is also zero, indicating that more than half of the organizations had no trained personnel in 2017. The 75th percentile is 10, indicating that the top 25% of organizations sent at least ten people to training. The 90th percentile climbs to 40, and the 95th percentile jumps to 88, suggesting that, while most companies trained very few people, a small percentage of organizations trained much more employees. The highest recorded number is 3,527, indicating that a corporation has the most

[22] The graph is obtained with:
`histogram vB11, title("Trained employees distribution (2017)") color(black) percent`

personnel trained. The mean of 22.61 is relatively low, however this is due to the large number of zero values in the sample.

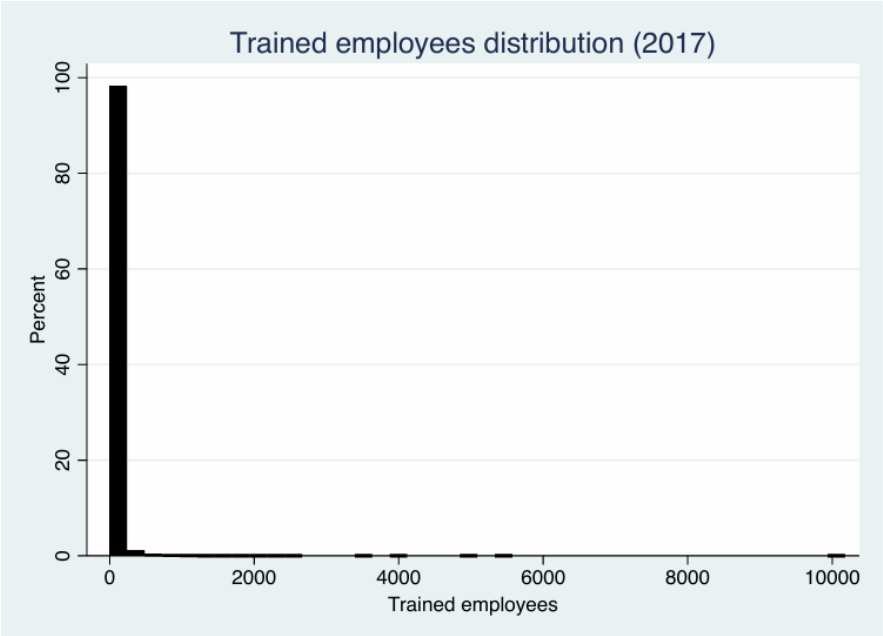

Figure 3: trained employees distribution

The standard deviation is relatively considerable (138.09), indicating a significant variation in the number of people trained between organizations. This implies that a few organizations have many trained staff, whilst the majority have few or none at all. The variance of 19,067.79 is similarly considerable, indicating a widespread in the data. The skewness of 31.20 shows a considerable rightward skew, as expected given the large number of zero values and a few enterprises with significantly more trained staff.

The kurtosis of 1,624.04 suggests a highly peaked distribution with heavy tails, implying that there are many extreme outliers with far more trained staff than most organizations.

To summarize, the distribution of `vB11` is substantially skewed, with many organizations reporting no people trained and only a few companies educating many employees. The strong skewness and kurtosis indicate that a few organizations stand out with unusually high training rates.

Training expense

Another of the main variables from the original dataset is now given. The variable `vB15_2` tracks the costs incurred by businesses for on-the-job training. At first scan, the variable displays 9,131 observations (see *Code.4* in the appendix). While it is likely that somewhat more than 30% of organizations did not record their training expenses (given that only 12,237 enterprises, or almost 40% of those polled, reported that they provided training) the number of observations for `vB15_2` appears to be low.

To enhance the number of relevant observations for subsequent analyses, the value of `vB15_2` was adjusted to zero when missing for organizations that stated they provided training. To accomplish this, the following command was executed:

```
replace vB15_2 = 0 if vB10 == 1 & missing(vB15_2)
```

To obtain enough observations for the purposes of this thesis, the data manipulation described above was deemed appropriate. It is reasonable to expect that for mandated training initiatives or training delivered through shadowing, businesses may state that they did not spend any money on training. At the same time, this assumption is controversial, as one could argue that time spent monitoring or conducting required training detracts from output creation and should thus be recorded as a training cost. However, it is important to note that providing an exact cost estimate for this form of training is especially difficult because such expenses are not usually precisely or regularly recorded.

In *Table 10* the variable `vB15_2` is presented: it represents how much companies spend on on-the-job training, with a total of 12,237 observations. As previously stated, we replaced missing numbers with zeros for companies who claimed providing training but did not identify any specific expenses. This change, while required to improve the number of useable observations, has unavoidably pushed the distribution closer to zero, increasing the already high concentration of low values. Looking at the distribution, the 1st, 5th, 10th, and 25th percentiles are all zero, indicating that at least a quarter of the organizations either did not spend formal training costs or did not explicitly account for them.

Ammontare della spesa per la formazione del personale nel corso del 2017

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	12,237
25%	0	0	Sum of Wgt.	12,237
50%	1200		Mean	57629.35
		Largest	Std. Dev.	2110555
75%	6000	3.29e+07		
90%	25000	7.37e+07	Variance	4.45e+12
95%	54468	8.21e+07	Skewness	78.1934
99%	300000	2.00e+08	Kurtosis	6896.401

Table 10: variable vB15_2 description

This finding supports the theory that certain businesses, particularly those that offer required or unofficial training (such mentorship or shadowing), might not record these costs as actual expenses. At 1,200 euros, the median (50th percentile) indicates that half of the businesses invested at least this much in training. In the higher percentiles, the distribution gets more diversified: the 95th percentile

rises to 54,468 euros, the 90th percentile reaches 25,000 euros, and the 75th percentile reaches 6,000 euros. The 99th percentile, which reaches 300,000 euros, reveals a select few businesses that spend far more on training. Extreme outliers have a significant impact on the mean training expense of 57,629.35 euros. *Figure 4* shows graphically the distribution[23]. The incredibly high variance of 4.45e+12 and the standard deviation of 2,110,555 euros make this clear. The distribution's skewness value of 78.19 indicates a noticeable right skew. This extreme skewness indicates that a few numbers of organizations devote significant money to training, resulting in a long tail to the right, while the majority report little to no investment. The kurtosis of 6,896.40, which shows a strongly peaked distribution with thick tails—a classic indication of outliers—further supports this pattern. This distribution's form is in line with what is sometimes referred to in the literature as a "zero-inflated distribution" (see Cameron & Trivedi, 2013), which happens when a variable has a lengthy right tail and an excess of zero observations. This pattern is especially prevalent in data pertaining to business investments, when a handful of enterprises report making disproportionately large investments while the majority record no spending at all. In conclusion, because of the data imputation procedure and the fact that many businesses do not disclose training costs directly, `vB15_2` is strongly concentrated around zero. Extreme dispersion results from a tiny percentage of businesses reporting extremely high expenditures.

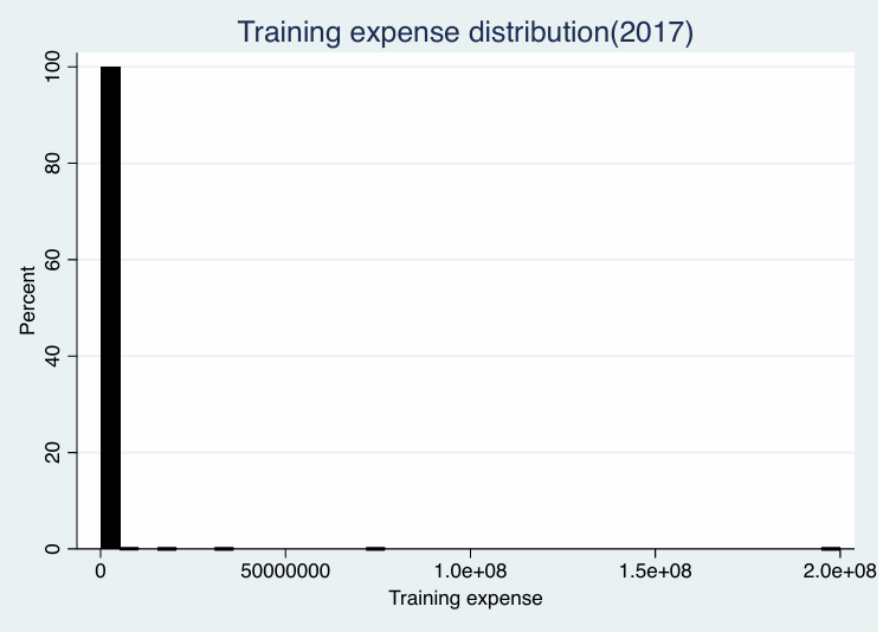

Figure 4: training expense distribution

[23] The graph is obtained with:

```
histogram vB15_2, title("Training expense distribution (2017)") color(black)
percent
```

Creation of productivity variable

The development of new key variables, which are based on those that were previously introduced and covered in the *Selection and presentation of key variables* section, is the main objective of this chapter. To provide a more thorough examination of the dataset, this stage aims to produce extra indicators that capture more subtle facets of business behaviour and staff training procedures.

Since it enables a more thorough and accurate examination of the connections among employee training, business attributes, and performance results, the development of these variables is an essential step in getting the data ready for later econometric modelling.

Productivity

The first custom variable created for this analysis is productivity, calculated as the ratio of company revenue (`vH13`) to total hours worked (`vH12`)[24]:

```
generate productivity_hours = vH13/vH12
```

This variable gives a clearer picture of how well businesses use their workers by comparing output to labor input, which helps to measure company performance more accurately.

However, potential data inconsistencies must be addressed before creating this variable. In particular, the need is to make sure that the amount of work that each employee does not above the legal cap in Italy, which is 48 hours per week for 52 weeks of the year, or a total of 2,496 hours per year. Observations violating this threshold likely result from reporting errors or outliers and must be excluded to maintain data reliability. To avoid mistakenly dropping companies with no employees (where `_v1=0`), the following conditional Stata command was used:

```
drop if vH12 / _v1 > 2496 & _v1>0
```

This preserves the observations for businesses without employees by guaranteeing that only cases with a positive number of employees are examined for excessive working hours. This procedure resulted in the deletion of 6,812 observations (see *Code.6* in appendix). To preserve data dependability and guarantee that the `productivity_hours` variable represents actual and legally compliant labor conditions, a considerable sample size decrease was required. After this modification, the `productivity_hours` variable turns into a reliable gauge of firm-level effectiveness and is an essential part of the econometric research that follows.

[24] See *Code.5* in the appendix

A descriptive examination, showed in *Table 11* of the `productivity_hours` variable after it has been constructed provides important information about how it is distributed. After removing irrational numbers, such as instances in which the number of hours worked per employee over the 2,496-hour annual legal limit, the dataset has 13,334 observations.

The productivity distribution is significantly skewed, as displayed in *Figure 5*[25], with a mean of 5,991.16 and a median of 103.65. As further evidenced by the 99th percentile hitting 12,854.61 and the maximum value skyrocketing to 17.8 million, the difference between the mean and the median suggests the existence of severe outliers.

With the first, fifth, and tenth percentiles at 2.01, 24.35, and 36.18, respectively, the lower percentiles demonstrate that a sizable percentage of businesses have comparatively low production. With the 90th percentile at 554.75, and the 75th at 214.59, the distribution begins to broaden near the higher end. The notion of a heavy-tailed distribution is supported by the incredibly large standard deviation of 228,593.7. The kurtosis of 4,460.85 and the skewness of 63.31, which both imply that a tiny percentage of businesses have exceptionally high productivity levels, pushing the average much over the median, further confirm this.

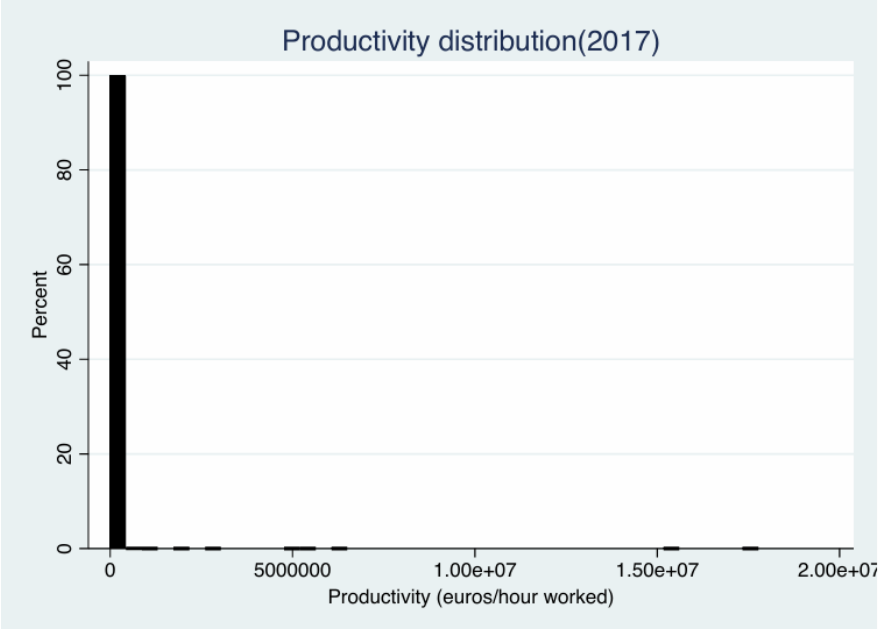

Figure 5: productivity distribution

The high number of zeros and small values, which are probably from businesses with little to no employees, along with the strong right skewness, indicate that any econometric modelling including this variable would profit from a robust regression technique or a logarithmic transformation.

[25] The graph is obtained with:

```
histogram productivity_hours, title("Productivity distribution (2017)")
color(black) percent
```

productivity_hours

	Percentiles	Smallest		
1%	2.014711	.0056423		
5%	24.35065	.0127955		
10%	36.18186	.0193821	Obs	13,334
25%	59.12353	.0216667	Sum of Wgt.	13,334
50%	103.6504		Mean	5991.159
		Largest	Std. Dev.	228593.7
75%	214.5886	5250571		
90%	554.7487	6250000	Variance	5.23e+10
95%	1185.963	1.56e+07	Skewness	63.31064
99%	12854.61	1.78e+07	Kurtosis	4460.85

Table 11: Variable productivity_hours description

In the context of the econometric model driving the whole thesis, at equation (33), the `productivity_hours` variable plays a central role as the dependent variable, providing the key measure of firm efficiency. Understanding its distribution is crucial, as any unaddressed skewness or outliers could distort the relationship between training, company size, and employee education levels and their effect on productivity.

It's also critical to remember that the sample being analysed is quite diverse, comprising businesses from a wide range of industries, each with radically different resources, business models, and organizational structures. The dataset is enhanced by this diversity, but it also adds complexity, which emphasizes the need for cautious statistical handling to guarantee that the associations found in the model are not skewed by extreme values or sector-specific outliers.

Before beginning the regression analysis, it is required to create a collection of variables that will be utilized in the estimates. These variables are designed to capture important features of workplace training and its possible impact on productivity. The following variables will be created.

The variable `diversification` of training indicates if a corporation provides at least one of four types of workplace training: mentoring, job-specific training, IT training, or other unspecified training. Mandatory training (i.e. `vB12_M_2`) is omitted since it does not serve as a differentiator between organizations, however the "other" category is preserved due to its general character and lack of precise classification. This operation is done thinking to the diverse utility of general and specific training explained by Becker (1964).

```
gen diversification = (vB12_M_1 == 1) + (vB12_M_3 == 1) + (vB12_M_4 == 1) +
(vB12_M_5 == 1)
```

The metric expressing the percentage of employees who have undergone training compared to the total number of employees is

```
gen trained_quota100 = (vB11 / _v1) * 100
```

Then, also the training expenditure per employee variable is considered, as it determines the total amount spent on training divided by the total number of employees.

```
gen expense_per_empl = vB15_2 / _v1
```

In addition, to assist the use of sector categorization in the study, the ATECO2007 variable will be translated into numerical format using the following command:

```
destring ATECO2007, replace
```

After creating these variables, a descriptive analysis will be performed to look at their distributions and detect any potential data inconsistencies. This will be done with the summary command:

```
sum diversification, detail
sum trained_quota100, detail
sum expense_per_empl, detail
sum ATECO2007, detail
```

The output of the summarization is reported in the appendices in *Table A.2*, *Table A.3*, *Table A.4* and *Table A.5*. Note that nothing strange appears to the dataset using the commands above.

5.2 RESULTS

The focus moves to regression models designed to investigate the relationship between workplace training and corporate productivity. The purpose of this section is to determine whether and to what degree investments in training affect business productivity levels.

The analysis will thus focus on confirming the existence of a statistically significant association between worker training and company productivity, as well as assessing the significance and robustness of the results considering the previously described methodological constraints.

Regression analysis for training and productivity.

In the following section, the regression analysis is presented to investigate potential correlations between productivity and various training-related variables. The study focuses on the link between the following variables: training expenditure, total number of trained employees, training percentage,

and training expenditure per employee. The study assumes that at least one form of training (mentoring, job-specific training, IT training, or other) has been offered, as indicated by the variable `diversification` greater than 1.

```
reg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1
```

Results are presented in *Table 12*.

Source	SS	df	MS	Number of obs	=	5,489
				F(4, 5484)	=	0.73
Model	3.9248e+10	4	9.8121e+09	Prob > F	=	0.5701
Residual	7.3528e+13	5,484	1.3408e+10	R-squared	=	0.0005
				Adj R-squared	=	-0.0002
Total	7.3567e+13	5,488	1.3405e+10	Root MSE	=	1.2e+05

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0000632	.0005457	-0.12	0.908	-.001133	.0010066
vB11	-.3121373	6.246535	-0.05	0.960	-12.55782	11.93355
trained_quota100	-82.34023	48.59039	-1.69	0.090	-177.5967	12.9162
expense_per_empl	.0210677	.1218693	0.17	0.863	-.2178444	.2599798
_cons	9956.501	3664.382	2.72	0.007	2772.858	17140.14

Table 12: regression results

The findings show that, at standard significance levels (e.g., 0.05), none of the major explanatory variables have a statistically significant impact on productivity. This statement can be looked at *Figure 6*, *Figure 7*, *Figure 8* and *Figure 9*[26].

With a p-value of 0.908, the training expense coefficient (`vB15_2`) is specifically negative but negligible, indicating that there is no meaningful correlation between training cost and productivity in this model. With a p-value of 0.960, the coefficient for the number of trained employees (`vB11`) is also modest and negative, further suggesting that there is no meaningful relationship between productivity and the number of trained personnel. The p-value of 0.090 suggests that this finding is only marginally significant at the 0.10 level, even though the coefficient for the percentage of trained staff (`trained_quota100`) is negative (-82.34).

[26] The graphs are obtained respectively with:

```
scatter productivity_hours vB15_2 if diversification>=1, title("Productivity vs. Training expense") msize(small) mcolor(black)

scatter productivity_hours vB11 if diversification>=1, title("Productivity vs. Trained employees") msize(small) mcolor(black)

scatter productivity_hours trained_quota100 if diversification>=1, title("Productivity vs. Trained percentage") msize(small) mcolor(black)

scatter productivity_hours expense_per_empl if diversification>=1, title("Productivity vs. Training expense per employee") msize(small) mcolor(black)
```

Figure 6: productivity vs. training expense

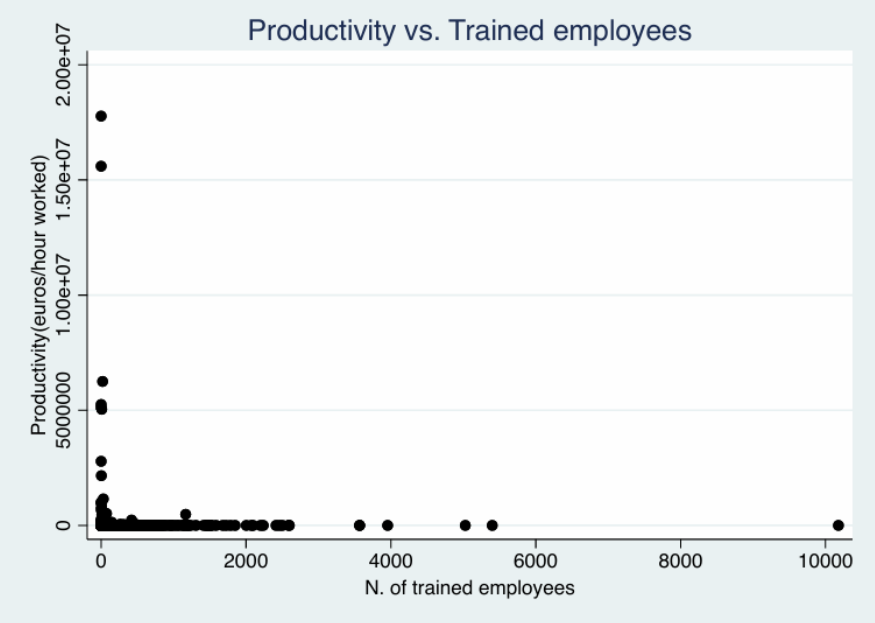

Figure 7: productivity vs. trained employees

However, with a p-value of 0.863, the coefficient for training cost per employee (`expense_per_empl`) is positive but negligible, indicating no meaningful correlation with productivity. Together with these side effects, the model's extremely low R-squared value of 0.0005 indicates that the variables only partially account for the variation in productivity. At 0.73, the overall F-statistic is likewise low, suggesting that the model does not adequately match the data.

Given the outcomes of the previous regression analysis, it is essential to conduct further tests with the aim to ascertain the appropriateness of implementing robust regression (`rreg`) in place of traditional OLS.

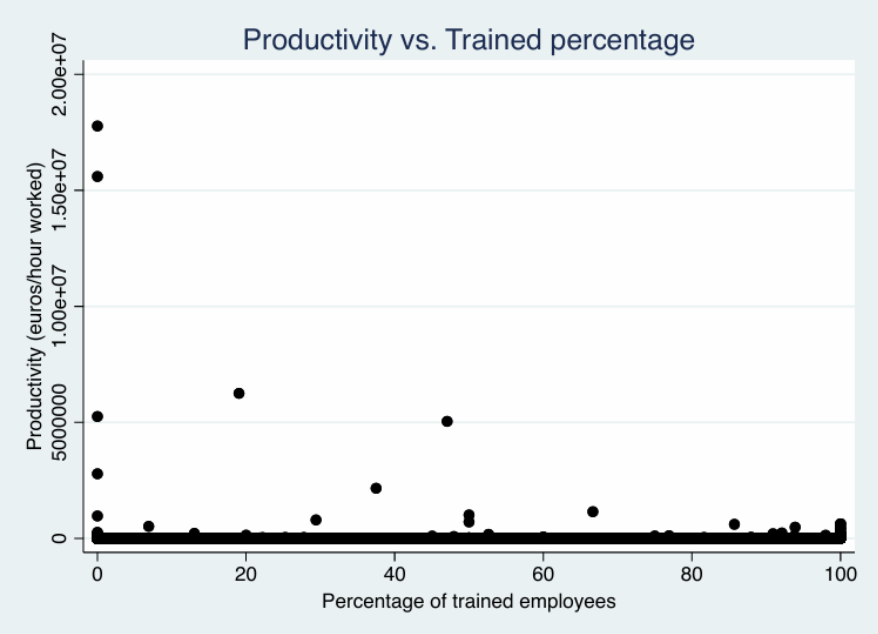

Figure 8: productivity vs. trained percentage

The presence of heteroscedasticity and influential outliers within the dataset leads to distortion in the results, thereby necessitating the resolution of these issues to yield more reliable estimates. The following describes the tests employed to evaluate the suitability of robust regression over OLS.

Figure 9: productivity vs. training expense per employee

Heteroscedasticity Test

One of the core premises of OLS regression is that the error terms exhibit homoscedasticity, meaning the variance of the residuals remains consistent across various levels of the independent variables. To identify the presence of heteroscedasticity, researchers often utilize the Breusch-Pagan test. If the Breusch-Pagan test indicates heteroscedasticity (which is common in cross-sectional data like this),

then the standard errors from the OLS regression would be unreliable, potentially leading to biased inferences about the significance of the coefficients.
The command used to this aim is

```
estat hettest
```

and the results are shown in *Figure 10*.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of productivity_hours

         chi2(1)      =  3407.22
         Prob > chi2  =   0.0000
```

Figure 10: output of the Het-test

The result of the test for heteroskedasticity shows a chi-squared value of 3407.22 and a p-value of 0.0000. The extremely low p-value provides strong evidence against the null hypothesis[27]. Hence, the null hypothesis is declined, and it can be said that heteroskedasticity is present in the model, indicating that the error variance is not constant across observations.
It is essential to keep in mind that outliers are not the result of some errors. These outliers are an expected outcome due to the diversity of the sample, which includes very small companies to large ones. The large range of firm sizes in the sample clearly results in some extreme statistics. The regression model must account for these outliers even though they have little bearing on the analysis. Heteroskedasticity increases the likelihood that the standard errors are not accurate, which may affect statistical findings such as hypothesis testing. Because robust regression adjusts for heteroskedasticity and ensures more accurate estimates by considering the impact of outliers and the unequal variance of errors, it is justified in this regard.

Influence of outliers

The existence of significant outliers, which might disproportionately impact the outcomes of an OLS regression, is another significant issue in regression analysis. The Cook's distance metric was employed, measuring the effect of individual data points on the calculated coefficients and checking for the effects of outliers. Regression results may be skewed by significant outliers, which are shown by points with high Cook's distance values. To determine and show Cook's distance, use the Stata command

[27] The null hypothesis states that the error variance is constant (homoscedasticity).

```
predict cooksd, cooksd
```

Next, to identify potential outliers, we can display the Cook's distance values. Points that are deemed influential and may require additional research are those whose Cook's distance is larger than 4 divided by the number of observations (_N). Some organizations may report excessive levels of productivity and training investment because the dataset in this study includes firms with a range of sizes and features. These findings might not be mistakes but rather represent real differences between businesses. Robust regression is favoured in these situations because it reduces the impact of these extreme values without eliminating them, enabling more precise estimations.
So, the command useful to show the outliers (considering that Stata considers missing values as greater than any value), and the relative output (as *Figure 11*) are presented below.

```
list cooksd if cooksd > 4/(_N) & cooksd != .
```

	cooksd
177.	.0037468
800.	.0005406
3620.	.0018353
3955.	.0024329
4891.	.0009939
4955.	.0002474
5177.	.0032607
5310.	.0004585
5931.	.001031
5968.	.0009271
6036.	.1119607
7342.	.0245337
8972.	.0002791
9157.	.001158
9321.	.0002987
9833.	.0021855
13275.	.0036498
14018.	.0021551
14108.	.004179
15162.	.0128241
15174.	.0042092
15675.	.0013399
16949.	.000485
17138.	.3539491

Figure 11: outlier representation

Values for several observations where Cook's distance over the threshold are provided by the Cook's distance list output. There are a few extreme values, such as the observation with an exceptionally high Cook's distance of 0.3539491[28], even if most of the values are somewhat tiny. This suggests that the regression results are disproportionately impacted by this specific observation[29].

[28] Observation number 17138.
[29] Even if already said, it is important to remind that the wide range of company sizes in the sample leads to some extreme numbers.

It is essential to take these significant observations into consideration in the analysis due to their existence. This provides additional support for the use of robust regression, which lessens the undue influence of extreme observations and guarantees more accurate coefficient estimates.
When heteroscedasticity and the impact of outliers are both issues, as they are in the current investigation, robust regression is especially helpful. Robust regression accounts for these problems by using a weighted method that reduces the impact of outliers, in contrast to Ordinary Least Squares (OLS), which assumes homoscedasticity and is sensitive to extreme observations. Heteroscedasticity, which shows that the error variance varies among observations, was found in this study using the Breusch-Pagan test. A range of observations from very small to extremely large organizations are also included in the dataset, which leads to some significant outliers. These outliers reflect the sample's inherent variability rather than data entry problems. To ensure that the regression results are not disproportionately impacted by extreme values, Stata's rreg technique uses an iterative approach to assign lower weights to these significant data. This method is the most suitable for this analysis since it yields more accurate coefficient estimations in the presence of heteroscedasticity and non-normal residuals.

Analysis of training and productivity correlation trough a robust regression.

Thus, it is now appropriate to request that Stata perform the robust regression.

```
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1
```

This Stata function handles possible problems like heteroscedasticity and outliers while doing a robust regression to investigate the connection between investments in on-the-job training and productivity. The variable `productivity_hours`, which calculates productivity in terms of hours worked, is the dependent variable in this regression. The independent variables are `trained_quota100`, which measures the proportion of trained employees to the total workforce, `vB11`, which shows the absolute number of trained employees, `expense_per_empl`, which records the training expenditure per employee, and `vB15_2`, which represents the total expenditure on training. Only companies that have participated in at least one kind of optional training—such as IT training, specialized training, mentorship, or other undefined forms—will be considered in the study thanks to the requirement `if diversification>=1`. Robust regression, in contrast to normal OLS regression, uses an iterative weighting method that lessens the impact of extreme values. This makes

it especially helpful for datasets where heteroscedasticity and outliers, as demonstrated by earlier diagnostic tests, could otherwise skew the findings.

In *Figure 12* the output of the robust regression is presented, while in *Figure 13* there is the "fitted vs. residuals" graph[30].

```
   Huber iteration 1:  maximum difference in weights = .99945951
   Huber iteration 2:  maximum difference in weights = .98282161
   Huber iteration 3:  maximum difference in weights = .84884849
   Huber iteration 4:  maximum difference in weights = .42489971
   Huber iteration 5:  maximum difference in weights = .13554278
   Huber iteration 6:  maximum difference in weights = .05777062
   Huber iteration 7:  maximum difference in weights = .0295481
Biweight iteration 8:  maximum difference in weights = .29356888
Biweight iteration 9:  maximum difference in weights = .06738103
Biweight iteration 10:  maximum difference in weights = .07817764
Biweight iteration 11:  maximum difference in weights = .05481277
Biweight iteration 12:  maximum difference in weights = .03173645
Biweight iteration 13:  maximum difference in weights = .01605809
Biweight iteration 14:  maximum difference in weights = .00807197

Robust regression                               Number of obs     =      5,489
                                                F(  4,      5484) =       6.06
                                                Prob > F          =     0.0001
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	5.92e-08	4.06e-07	0.15	0.884	-7.37e-07	8.55e-07
vB11	.0188265	.0046476	4.05	0.000	.0097153	.0279376
trained_quota100	-.1111183	.0361527	-3.07	0.002	-.1819919	-.0402446
expense_per_empl	9.68e-06	.0000907	0.11	0.915	-.0001681	.0001874
_cons	115.3439	2.72641	42.31	0.000	109.999	120.6887

Figure 12: robust regression output

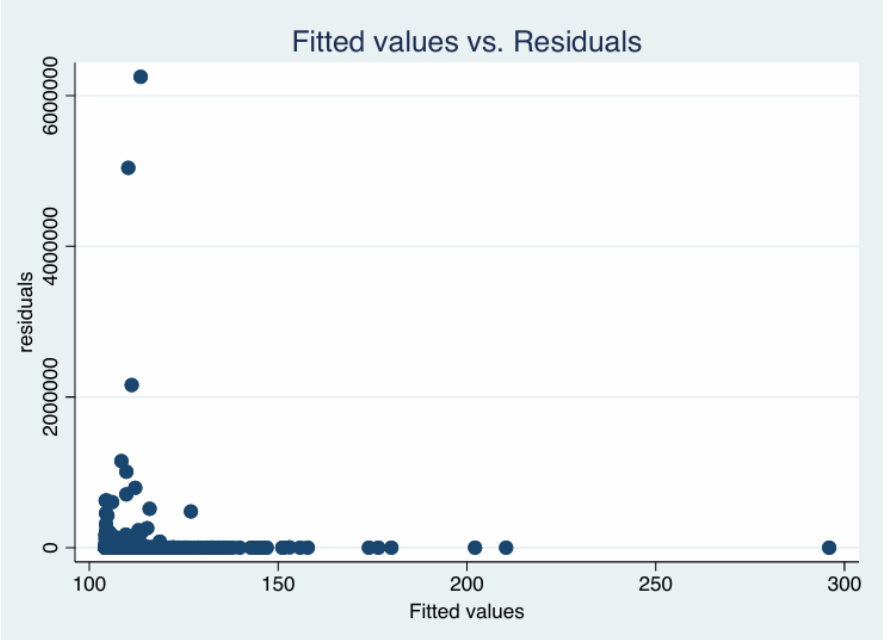

Figure 13: fitted values vs. residuals

[30] The graph is obtained with:

```
predict fitted_values, xb
predict residuals, residuals
scatter residuals fitted_values, title("Fitted values vs. Residuals) mcolor(black)
```

A significant alternate viewpoint on the relationship between training costs, the number of employees trained, the percentage of trained employees, and production is offered by the robust regression results. The analysis takes into consideration possible outliers and heteroscedasticity, which were problems found in the previous OLS regression, by using the robust regression method. This method reduces the effect of extreme values and yields more accurate estimations, especially when there are data irregularities.

Initially, we note that the coefficient for `vB15_2`, which stands for the overall training cost, is a very modest and positive value of 5.92e-08, with a corresponding p-value of 0.884. This implies that there is no discernible effect of training cost on production, in contrast to the findings of the OLS regression, which found no statistically significant association between training expense and output. The extremely low coefficient demonstrates that any impact of training costs on productivity is minimal. This is in line with previous findings that showed a very weak and statistically negligible link between production and training cost, as determined by OLS regression.

It can be observed a more substantial outcome when moving to the coefficient for `vB11`, which stands for the absolute number of trained employees. There is a statistically significant positive correlation among production and the number of trained employees, as indicated by the coefficient of 0.0188 and the p-value of 0.000. The robust regression makes it obvious that having more trained personnel has an advantageous impact on productivity, which is in line with the results of the OLS regression, which also showed a very modest negative coefficient for the number of trained employees. Therefore, in contrast to the OLS regression that yielded a non-significant outcome, robust regression analysis indicates that workforce training enhances productivity, and this finding holds statistical significance.

The coefficient for the variable `trained_quota100`, which indicates the percentage of trained personnel compared to the overall workforce, is -0.1111, with a p-value of 0.002, suggesting a significant negative connection with productivity. This shows that increasing employee training as a percentage of the whole workforce may reduce productivity. This finding contrasts with the OLS regression results, which showed that the percentage of trained staff had a modestly negative influence (statistically significant at the 0.10 level). However, the robust regression shows a greater, statistically significant negative association, implying that the percentage of skilled staff, rather than the overall number, may be a more important factor in determining productivity. This could imply that, while increasing the number of skilled people is desirable, investing excessively in training a big proportion of the workforce may not result in the same productivity gains.

With a value of 9.68e-06 and a p-value of 0.915, the coefficient for `expense_per_empl`, which calculates the training expenditure per employee, is incredibly small, indicating that there is no statistically significant correlation between productivity and training expenditure per employee. The

OLS regression results, which likewise showed a modest and negligible coefficient for training spend per employee, are consistent with this finding. The training spend per employee has no discernible impact on production, even when the robust model accounts for heteroscedasticity and outliers. According to both regression models, this implies that the way training expenses are allocated among staff members has no discernible effect on total production.
When all explanatory factors are zero, the baseline productivity value is statistically significant, as indicated by the constant term (`_cons`) of 115.344 and its p-value of 0.000. In the absence of training costs, trained personnel, or any other type of workforce development investment, this is the expected productivity. The big coefficient, which is highly statistically significant, represents the sample's firms' overall productivity level before taking into consideration the training-related variables.

In conclusion, several significant insights are provided by the robust regression results. First, the number of trained employees (`vB11`) is now statistically significant and positively correlated with productivity, indicating that raising the absolute number of trained employees boosts productivity, even though the effect of training expenditure (`vB15_2`) on productivity is still insignificant in both the OLS and robust regressions. Productivity is significantly impacted negatively by the `trained_quota100` percentage of employees, suggesting that there may be a declining return on investment in raising the percentage of trained personnel. Lastly, neither regression model demonstrates a significant correlation between productivity and the spending per employee (`expense_per_empl`), supporting the notion that production is not greatly impacted by the distribution of training expenditures across employees.

Given the existence of outliers and heteroscedasticity, which were major issues with the OLS model, the findings are that the robust regression has produced more dependable and consistent estimates when comparing these results to the previous OLS regression. By accounting for these problems, the robust regression approach offers a more accurate representation of the actual correlations between the variables. Given the characteristics of the data, robust regression was selected as a more suitable method for this study. This is further supported by the statistical significance of the number of trained workers and the negative association with the percentage of employees trained.

Sectors for which there is statistical significance.

Firstly, economic sectors that present a causal relation for on-the-job training and productivity were presented; then, sectors with no sufficient data or weak correlations are presented. Finally, the focus will be on sectors that do not show any effect of OJT training on productivity.

The manufacturing sector

Since the manufacturing sector is the main one mentioned in the data sources that are available, the first focus of the analysis that follows moves to this industry. The `ATECO2007` categorization is used to identify the manufacturing sector, corresponding to the codes from 10 to 33. The robust regression model is executed using the following command to separate this subset of the data:

```
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=10 & ATECO2007<=33
```

This directive limits the analysis to manufacturing firms that have adopted at least one training program (as shown by `diversification >= 1`). By using this filter, the focus is reduced to a more homogeneous and specific set of businesses, making it easier to comprehend the relationship between productivity in the manufacturing sector and training expenditures, employee numbers, employee percentages, and training expenditures per employee.

After the output is accessible, the analysis of the findings is accessible, and it is possible to determine whether the connections found in the broad study still apply to this industry.

After proposing to Stata the command described, the output get is the one in *Figure 14*.

```
   Huber iteration 1:  maximum difference in weights = .99887746
   Huber iteration 2:  maximum difference in weights = .99058981
   Huber iteration 3:  maximum difference in weights = .64200321
   Huber iteration 4:  maximum difference in weights = .34901842
   Huber iteration 5:  maximum difference in weights = .09706578
   Huber iteration 6:  maximum difference in weights = .03196329
Biweight iteration 7:  maximum difference in weights = .29401323
Biweight iteration 8:  maximum difference in weights = .10954202
Biweight iteration 9:  maximum difference in weights = .03087762
Biweight iteration 10:  maximum difference in weights = .01354201
Biweight iteration 11:  maximum difference in weights = .00520406

Robust regression                               Number of obs     =      2,083
                                                F(  4,      2078) =      56.42
                                                Prob > F          =     0.0000
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-4.62e-07	4.86e-07	-0.95	0.341	-1.41e-06	4.90e-07
vB11	.1047182	.0070054	14.95	0.000	.0909799	.1184565
trained_quota100	-.0619756	.0590283	-1.05	0.294	-.1777365	.0537852
expense_per_empl	.0004313	.0003708	1.16	0.245	-.0002959	.0011585
_cons	124.2499	4.181246	29.72	0.000	116.05	132.4498

Figure 14: robust regression results for the manufacturing sector

When comparing the robust regression output for the manufacturing sector to the general regression output, some significant differences are evident. The manufacturing sector sample has 2,083 observations, which is significantly fewer than the 5,489 in the general dataset, even though it is almost half of that. This restriction results in different findings, especially when it comes to the

relationships between the explanatory variables and productivity. One significant discrepancy lies in the coefficient for training expenditure (vB15_2), which is -4.62e-07 within the context of the manufacturing regression. This coefficient is remarkably minuscule and negative, boasting a p-value of 0.341. This suggests that there is no substantial relationship to be discerned between training expenditure and productivity within the domain of the manufacturing sector. When examining the broader, more general regression, an analogous outcome emerges, where the vB15_2 coefficient is observed to be equally small and negative, though its significance diminishes further, as indicated by an elevated p-value of 0.908. The lack of statistical significance in both situations suggests that, despite its negative magnitude, training spending has no discernible impact on either sector's productivity. It's possible that other elements, including business-specific tactics or outside market circumstances (as Dearden et al. (2006) said), have a greater impact on productivity than training costs alone.

The manufacturing sector's regression shows a positive and highly significant coefficient of 0.1047 with a p-value of 0.000 for the number of trained personnel (vB11). This implies that productivity in the industrial sector is significantly positively impacted by the quantity of trained workers. This outcome contrasts with the general regression, which presented a negative (-0.3121) yet non-significant coefficient for vB11. The outstanding positive correlation observed within the manufacturing sector underscores the significance of an increased number of skilled workers for enhancing productivity in that specific area. It is plausible that a greater number of skilled individuals can lead to an immediate enhancement of production levels in manufacturing, a domain where specialized knowledge and competencies are frequently indispensable.

The manufacturing sector regression's coefficient for the percentage of trained workers (trained_quota100) is -0.06198, which is negative and not statistically significant (p-value = 0.294). In contrast, the general regression showed a meaningfully negative coefficient (-82.34) that was only marginally significant (p-value = 0.090). The manufacturing regression's lack of significance suggests that there is no discernible or significant effect of the overall proportion of trained personnel on productivity, as compared to the total number of trained employees. This may be because other variables affecting productivity in this industry are more prevalent, or because the quality or kind of training may be more significant in manufacturing than merely raising the proportion of skilled workers.

There is no discernible impact on productivity, as indicated by the manufacturing sector's training expenditure per employee (expense_per_empl) coefficient of 0.0004313 and p-value of 0.245. This outcome is in line with the general regression, which likewise showed a tiny and non-significant coefficient for expense_per_empl (p-value = 0.863). The absence of a significant correlation in both

situations indicates that neither the manufacturing sector nor the general sample's productivity is significantly influenced by the cost of training per employee.
Lastly, the manufacturing sector regression's constant term is 124.2499, significantly less than the overall regression's 9956.501 constant term. This discrepancy represents the manufacturing sector's baseline productivity, which is probably not the same as the entire dataset. Because of the size of the industry or the intrinsic structure of manufacturing processes, the lower constant value may suggest that the enterprises in this sector begin with a distinct level of productivity.

In summary, the manufacturing sector's regression results demonstrate that, in contrast to the general regression, the quantity of skilled workers significantly increases productivity. This discrepancy could result from the unique characteristics of manufacturing, where educating more workers could result in more significant productivity gains. This could be because industrial operations frequently call for specialized knowledge and abilities. Both the general and manufacturing regressions' lack of significance for training spending and training per employee, however, raises the possibility that other factors may be more important in explaining changes in productivity in these industries.

The construction sector

The results indicate that applying the robust regression model to the construction industry might provide insightful information, even though this industry is not specifically addressed in the data sources that are currently accessible. Thus, this sector is the focus of the analysis that follows. The construction industry is defined under the ATECO2007 classification, which corresponds to codes 41 to 43. To separate this subset of the data, the robust regression model is run using the following command:

```
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 > & ATECO2007>=41 & ATECO2007<=43
```

Due to this instruction, the research is limited to construction companies who have at least one specific training program in place, as already proposed and discussed. *Figure 15* displays the result that was produced when the command was executed in Stata. There are also significant variations between the manufacturing sector and general regressions and the robust regression findings, which are based on 692 observations. The stability of the estimations may be impacted by the much smaller amount of data.

```
   Huber iteration 1:  maximum difference in weights = .9956094
   Huber iteration 2:  maximum difference in weights = .9739058
   Huber iteration 3:  maximum difference in weights = .82354693
   Huber iteration 4:  maximum difference in weights = .91374274
   Huber iteration 5:  maximum difference in weights = .21598342
   Huber iteration 6:  maximum difference in weights = .03459078
Biweight iteration 7:  maximum difference in weights = .29209225
Biweight iteration 8:  maximum difference in weights = .09489982
Biweight iteration 9:  maximum difference in weights = .03800598
Biweight iteration 10:  maximum difference in weights = .00733299

Robust regression                               Number of obs     =        692
                                                F(  4,       687) =      37.96
                                                Prob > F          =     0.0000
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0004514	.0000802	-5.63	0.000	-.0006089	-.0002939
vB11	.4685936	.0421005	11.13	0.000	.3859325	.5512548
trained_quota100	-.1641675	.0630873	-2.60	0.009	-.2880345	-.0403004
expense_per_empl	.0207768	.0027458	7.57	0.000	.0153855	.026168
_cons	83.59289	4.777935	17.50	0.000	74.21178	92.974

Figure 15: robust regression results for the construction industry

With an F-statistic of 37.96 and a p-value of 0.0000, the model is still very significant and shows that the explanatory variables have a considerable impact on the explanation of production fluctuations.

With a coefficient of -0.0004514 and a p-value of 0.000, training expense (`vB15_2`) significantly reduces production, in contrast to the earlier regressions. In contrast to the general and manufacturing regressions, where the effect of training expenditure was statistically insignificant, this shows that more training investment is linked to poorer productivity in the industrial sector. This can be a sign of inefficient training expenditure allocation in this industry or that training causes production process disruptions, which momentarily lowers productivity.

Compared to the two earlier regressions, the number of trained personnel (`vB11`) has a significantly larger positive correlation with productivity, with a coefficient of 0.4686 and a p-value of 0.000. Given the technical nature of industrial labour, where specialized skills are crucial, this conclusion implies that increasing the absolute number of skilled people has a more significant impact on productivity in the construction sector.

With a p-value of 0.009 and a coefficient of -0.1642, the proportion of trained personnel is statistically significant. This adverse impact implies that a larger percentage of skilled workers does not always equate to increased productivity, which is different from the non-significant findings in the manufacturing regression. This might suggest that it is more beneficial to teach a select few important employees rather than dispersing training efforts widely.

With a coefficient of 0.0208 and a p-value of 0.000, training expenditure per employee (`expense_per_empl`) is substantially positive, in contrast to the preceding regressions. This implies that training expenditures have a beneficial influence on industrial sector productivity when seen as a per-employee expense rather than as an overall cost, supporting the notion that more focused, superior training is more successful.

All things considered, the findings imply that training expenditures and their outcomes differ greatly among industries. While total training investment seems to be counterproductive, the number of trained individuals has a large beneficial influence in the construction sector. These variations show that industry-specific training methods are required instead of a one-size-fits-all strategy.

The retail sector

Results from the retail trade sector, which includes also auto and motorcycle maintenance, suggest that using the robust regression model in this area might yield insightful information. This industry is identified by codes 45 to 47 in the ATECO2007 categorization, which includes operations pertaining to wholesale, retail, and automobile trade and repair.

Examining whether the links seen in other industries apply here is especially pertinent given the sector's economic importance and structural distinctions from manufacturing and construction. The analysis's findings, presented in *Figure 16* and obtainable with the command

```
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 > & ATECO2007>=45 & ATECO2007<=47
```

will be helpful to comprehend the function of training and human capital development in this industry and how it could affect company performance.

```
   Huber iteration 1:  maximum difference in weights = .99797692
   Huber iteration 2:  maximum difference in weights = .97200283
   Huber iteration 3:  maximum difference in weights = .77521029
   Huber iteration 4:  maximum difference in weights = .58280466
   Huber iteration 5:  maximum difference in weights = .17470488
   Huber iteration 6:  maximum difference in weights = .02152461
Biweight iteration 7:  maximum difference in weights = .2940915
Biweight iteration 8:  maximum difference in weights = .0554598
Biweight iteration 9:  maximum difference in weights = .01455003
Biweight iteration 10:  maximum difference in weights = .00587661

Robust regression                               Number of obs     =        625
                                                F(  4,       620) =       2.59
                                                Prob > F          =     0.0358

------------------------------------------------------------------------------
productivity_h~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
          vB15_2 |   .0000363   .0000161     2.26   0.024     4.71e-06    .0000679
            vB11 |   .0100895   .0228055     0.44   0.658    -.0346958    .0548748
trained_quota100 |   .3693617   .1778325     2.08   0.038     .0201346    .7185887
expense_per_empl |  -.0006391   .0003498    -1.83   0.068    -.0013261    .0000478
           _cons |   166.7752   12.73214    13.10   0.000     141.7719    191.7786
------------------------------------------------------------------------------
```

Figure 16: robust regression output for the retail sector

Based on 625 data, the strong regression results for the wholesale and retail trade as well as the motor vehicle and motorcycle maintenance industry reveal some clear trends in contrast to earlier assessments. The model's explanatory power is lower than that of the regressions performed on the

general sample, the manufacturing sector, and the construction sector, even though the F-statistic of 2.59 and the p-value of 0.0358 show that the model is statistically significant.

A remarkable distinction is that training expenditure (`vB15_2`) now has a positive impact on production and is statistically significant (coefficient = 0.0000363, p-value = 0.024). This contrasts with earlier research showing that training spending was either negligible or even inversely correlated with construction sector production. Given the service-oriented nature of the trade and repair sector, where abilities like product knowledge, customer service, and technical repair skills may directly translate into improved business performance, this result implies that higher training expenditures are linked to higher productivity.

In contrast to the industrial sector, where this variable had a considerable positive effect, the number of trained personnel (`vB11`) had no discernible influence on productivity (coefficient = 0.0101, p-value = 0.658). This implies that just increasing the number of trained workers in trade and repair companies does not always translate into increased productivity. This might be because productivity increases in this industry depend more on individual competence than on extensive training initiatives.

A greater proportion of trained workers within a firm is linked to improved production, as indicated by the positively significant percentage of trained employees (coefficient = 0.3694, p-value = 0.038). This conclusion contrasts with the manufacturing sector, where it was not significant, and the construction sector, where the influence was negative. This implies that a well-trained staff as a percentage of total employees might boost firm performance in trade and repair operations, perhaps through improved customer service quality and operational efficiency.

However, training cost per employee (`expense_per_empl`) is only slightly significant, with a negative coefficient (-0.0006391) and a p-value of 0.068. This implies that although total training spending is beneficial, increased training expenses per employee might not necessarily result in corresponding increases in productivity. This might suggest that training investments per employee are declining or that some businesses may be overspending on training initiatives that don't immediately improve operational effectiveness.

Out of all the regressions that have been done thus far, the constant term (166.7752) is the greatest, indicating that the baseline productivity in this sector is substantially greater than that of manufacturing and industry. This is probably because trade and repair companies are distinct from other types of organizations in that they use different metrics to assess productivity, and factors like client flow, sales volume, and service efficiency can have a greater impact on productivity than manufacturing output.

All things considered, these findings demonstrate how the correlation between production and training differs greatly among industries. Both total training spending and the percentage of trained workers appear to be more important in the trade and repair sector than in manufacturing and construction industry, where training initiatives do not necessarily result in visible productivity increases. This supports the notion that knowledge and abilities are critical in boosting corporate success in a customer-facing, service-oriented industry. To achieve cost-effectiveness, businesses must strategically manage their training expenses, as seen by the slightly negative effect of training costs per person.

Transport and warehouse

Businesses falling within ATECO2007 codes 49 to 53 (which cover transportation by land, sea, and air as well as warehousing and transportation support activities) are included in the study, and its analysis is also intriguing. Understanding the connection between labour characteristics, training investments, and productivity is especially important given the sector's crucial position in logistics and supply chain efficiency. The robust regression model that follows is used to separate this group of data:

```
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 > & ATECO2007>=49 & ATECO2007<=53
```

This method limits the study to transportation and storage companies that have at least one specific training program in place, as always done. After the output is produced, the findings will be examined to see if the connections found in the larger research apply to this industry. *Figure 17* displays the result that was produced when the command was run in Stata.

Based on 256 data, the regression shows some significant variations from earlier sectoral assessments. Although the model is statistically significant ($F = 6.56$, $p < 0.001$), it seems to have less explanatory power than in the manufacturing sector. This is probably because the productivity factors in this business are different. One important conclusion is that productivity is positively and significantly impacted by training expenditure (`vB15_2`) (coefficient = 0.0002999, p-value = 0.004). This is consistent with findings from the extensive and retail trade sectors, where training expenditures also had a favourable impact. Investment in training seems to increase productivity in the transportation industry, in contrast to manufacturing, where training spending was not significant. This might be because efficiency and service quality are directly impacted by skill development in logistics, vehicle operation, and safety regulations.

```
   Huber iteration 1:  maximum difference in weights = .98371417
   Huber iteration 2:  maximum difference in weights = .94833687
   Huber iteration 3:  maximum difference in weights = .68152725
   Huber iteration 4:  maximum difference in weights = .49757612
   Huber iteration 5:  maximum difference in weights = .22406569
   Huber iteration 6:  maximum difference in weights = .12479773
   Huber iteration 7:  maximum difference in weights = .07411921
   Huber iteration 8:  maximum difference in weights = .0271376
Biweight iteration 9:  maximum difference in weights = .29019577
Biweight iteration 10:  maximum difference in weights = .17281059
Biweight iteration 11:  maximum difference in weights = .22800046
Biweight iteration 12:  maximum difference in weights = .16588401
Biweight iteration 13:  maximum difference in weights = .06413121
Biweight iteration 14:  maximum difference in weights = .01775256
Biweight iteration 15:  maximum difference in weights = .00470818

Robust regression                               Number of obs     =        256
                                                F(  4,       251) =       6.56
                                                Prob > F          =     0.0000

------------------------------------------------------------------------------
productivity_h~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
           vB15_2 |   .0002999   .0001039     2.89   0.004     .0000953    .0005046
             vB11 |   -.021107   .0130802    -1.61   0.108    -.0468679    .0046538
trained_quota100 |  -.2839566   .1809133    -1.57   0.118    -.6402581    .0723449
expense_per_empl |    .025397   .0090502     2.81   0.005      .007573     .043221
           _cons |   118.5624   13.44831     8.82   0.000     92.07646    145.0483
------------------------------------------------------------------------------
```

Figure 17: robust regression results for the transport and warehouse sector

On the other hand, merely increasing the number of trained personnel does not always result in increased output, as the number of trained employees (`vB11`) is not statistically significant (p = 0.108). In contrast to the construction sector, where the quantity of skilled workers had a significant positive impact, this is consistent with the findings from the trade sector. The outcome could suggest that the effectiveness of training in the transportation industry depends more on the calibre of trained personnel than on their number. In a similar vein, the `trained_quota100` has a negative coefficient (-0.284) and is not significant (p = 0.118).

In contrast to the trade sector, where a higher percentage of skilled employees had a beneficial impact on productivity, this is comparable to the construction sector, where a higher proportion of trained workers was linked to poorer productivity. One explanation might be that, rather than extensive training initiatives, technological advancements, infrastructure, and operational efficiency are what drive productivity improvements in the transportation and storage industries.

The amount spent on training per employee (`expense_per_empl`) has a substantial beneficial impact on production (coefficient = 0.0254, p = 0.005.005). This conclusion differs from other sectors in that. This is probably due to rigorous and excellent training programs that are aimed at enhancing operational effectiveness, legal compliance, and safety protocols.

According to the constant term (118.56), the sector's baseline productivity is comparable to manufacturing (124.25) but lower than trade (166.77). This outcome is consistent with the characteristics of the transportation sector, where external variables like infrastructure, fuel prices, and logistics networks impact productivity (according to Becker’s theory of externalities influencing productivity), while internal training expenditures are only one of several elements at play.

The results support the notion that there is a strong sector-dependent link between productivity and training as compared to earlier regressions. Although the quantity of skilled workers was important in the manufacturing sector, the efficiency of training expenditures seemed to be more important in the trade and transportation sectors. Productivity growths in storage and transportation, however, are more closely related to the extent and quality of training than to the percentage of personnel with training. This implies that rather than just raising training participation rates, businesses in this industry should place a higher priority on well-designed, highly effective training programs.

The artistic, sports and entertainment activities

The influence of employee training in a very heterogeneous business may be seen from the examination of productivity in the artistic, sports, and entertainment sector (ATECO 90–93). This industry includes both more conventional tasks like running sports facilities or entertainment venues as well as highly specialized occupations like artistic and creative labour.

In contrast to industries like manufacturing or transportation, where operational effectiveness and the number of skilled workers is key determinants of productivity, artistic and entertainment endeavours may rely on less obvious elements like the quality of training or creative service delivery. It is possible to determine if and how training investments affect productivity in companies operating in this sector by using a strong regression model. It is especially intriguing that there is a correlation between production and training even in this industry. This implies that staff training and skill development may still contribute to improving corporate performance, even in the face of the distinctive characteristics of the creative and entertainment sectors. Following the release of the regression output, the findings may be investigated and contrasted with those of other sectors that have already been studied. With this aim, Stata was asked to perform the robust regression in the following way:

```
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=90 & ATECO2007<=93
```

and the output returned is presented in *Figure 18*.

```
   Huber iteration 1:  maximum difference in weights = .76898683
   Huber iteration 2:  maximum difference in weights = .16295996
   Huber iteration 3:  maximum difference in weights = .13043402
   Huber iteration 4:  maximum difference in weights = .0022333
Biweight iteration 5:  maximum difference in weights = .07753737
Biweight iteration 6:  maximum difference in weights = .08583887
Biweight iteration 7:  maximum difference in weights = .01763299
Biweight iteration 8:  maximum difference in weights = .0090683

Robust regression                               Number of obs     =         12
                                                F(  4,          7) =      34.99
                                                Prob > F          =     0.0001
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0044719	.0012878	-3.47	0.010	-.0075171	-.0014266
vB11	3.347469	.3878941	8.63	0.000	2.430245	4.264692
trained_quota100	-.0761029	.4270762	-0.18	0.864	-1.085978	.9337717
expense_per_empl	.8610285	.394698	2.18	0.065	-.0722838	1.794341
_cons	25.59531	27.62516	0.93	0.385	-39.72781	90.91843

Figure 18:robust regression outout of the artistic, sports and entertainment sector

Particularly when compared to the earlier sectors examined, the robust regression's findings offer several intriguing insights.

First, despite the small sample size (12 observations), the model fits the data well, as seen by the very high F-statistic at 34.99 and the highly significant p-value (0.0001). The robustness of some of the results, however, could be constrained by the limited sample size. It is noteworthy that the training expenditure coefficient (`vB15_2`) is negative (-0.0044719) and statistically significant at the 1% level (p = 0.010) when compared to the coefficients from other sectors. In contrast, the impact of training expenditures was either insignificant or not statistically significant in many other areas.

In this instance, the negative coefficient implies that higher training costs may be linked to lower production for this industry. This finding may suggest that, because of the more ethereal and creative nature of the job, training initiatives may not necessarily result in quick or direct increases in productivity in industries like the arts and entertainment. It could also indicate a discrepancy between the kind of training offered and the abilities required in this industry.

The number of trained personnel (`vB11`) has a substantial positive coefficient (3.347469), and it is statistically significant (p = 0.000). This conclusion is consistent with findings from other industries where productivity and the number of trained workers is positively correlated. The idea that specialized abilities and talent are essential in the arts and entertainment industry is supported by the fact that this variable has such a big and positive influence in this sector, indicating that hiring more skilled workers greatly increases productivity. Similar to several other industries where the overall proportion of trained personnel did not have a significant influence, the percentage of trained employees had a very minor and statistically negligible effect on production (-0.0761029, p = 0.864). This implies that the depth or specialty of the training may be more important than the aggregate percentage of employees who have received training. As predicted, the coefficient associated with training investment per employee is positive (0.8610285) and only marginally significant at 10

percent ($p = 0.065$). While this is not as robust as the effects witnessed in other industries, it continues to suggest that there are aids to productivity in the creative industries when training, or skills development, is aimed at specific employees.

Compared to the results from areas like manufacturing or transportation that have a more direct connection between training and productivity, the findings for this sector are more subtle and less predictable. In the performing arts and entertainment segments, it seems that productivity is influenced by several additional factors (such as creativity, talent, and the type of training received). The negative coefficient for training expenditure, along with the substantial positive impact of the number of trained employees, indicates this. Hence, while this sector stands to gain from training, it is not a simple matter, and the outcome is likely more contingent on the context and training specifics than the sector itself.

Sectors with partial significative correlation.

Training and productivity have been found to be strongly correlated in some industries, but the situation is more complex in others. The association seems weak, inconsistent, or only partially supported by the data in several industries. The full regression outputs and code are provided in the appendix (from *Code.7* to *Code.9*) for reference, while the discussion below highlights the most relevant insights.

Only a small number of regressors show statistical significance, indicating restricted linkages in three sectors in particular: *mining and quarrying* (ATECO 5-9), *rental, travel agencies, and business support services* (ATECO 77-82), and *health and social work activities* (ATECO 86-88). These industries offer intriguing insights into the differing effects of training on productivity in various economic circumstances, despite not being the primary focus of the investigation.

According to the regression results, there is no statistically significant correlation between productivity and on-the-job training in the *mining* industry. Rather, company size (`vB11`), which exhibits a positive and substantial connection, seems to be the only pertinent element. This implies that larger businesses are typically more productive, maybe because of economies of scale, better machinery, or more organized operations. The capital-intensive character of the sector, where mechanization and technology investments have a greater impact on production than personnel training, may account for the lack of a relationship between training and output. Additionally, a lot of mining procedures call for official certifications or extremely specialized training prior to work, which may indicate that on-the-job training has less effect in comparison to prior knowledge.

A slightly different situation is presented by the *rental, travel agency, and business support services* sectors. Per employee training expenses exhibit a positive link with productivity, making them the

only regressor with a significant correlation. This implies that businesses that invest more per employee tend to be more productive. On-the-job training, however, shows no discernible association. One argument could be that operational simplification, client relations, and service efficiency (rather than formal training programs) are what increase productivity in this industry. Since many companies in this sector depend on automation, standard operating procedures, or customer-driven dynamics, experience and flexibility may be more important than formal training initiatives.

The regression results are especially interesting in the *healthcare* industry. There is a statistically substantial negative association between productivity and on-the-job training. This finding is unexpected and raises the possibility that lower productivity levels could be correlated with more training. One explanation could be that training frequently takes place in the healthcare and social work industries in response to skill shortfalls, regulatory revisions, or the requirement to onboard new employees, which momentarily lowers productivity. Furthermore, practical training might demand a lot of resources, including supervision and time away from providing direct patient care. In contrast to other sectors where training might boost productivity right away, the advantages of training in the healthcare industry might take longer to manifest, giving the impression that its short-term effects are detrimental.

These results demonstrate that although training is a major factor in increasing productivity across a wide range of businesses, its effects differ greatly based on sectoral characteristics. While service-based companies may gain more from other labor investments, such as increased pay, capital-intensive industries, like mining, may profit more from technology investments than from the growth of human capital. The healthcare industry's negative correlation raises the possibility that training's advantages could be momentarily outweighed by structural limitations or interruptions.

Sectors not analysable.

For some other sectors, data restrictions prevent a thorough assessment for several areas. Fields like *agriculture, forestry and fishing* (ATECO 01-03), *public administration and defense, compulsory social security* (ATECO 84), *households as employers of domestic personnel* (ATECO 97-98), and *extraterritorial organizations and bodies* (ATECO 99) either present datasets that are too fragmented or lack enough observations to support statistically significant conclusions. In the appendix, Stata code and outputs for these sectors are presented from (*Code.10* to *Code.13*, respectively).

In these situations, there may be a few reasons why there is insufficient data. For instance, compared to private-sector companies, public administration and defense have strict institutional frameworks that may not capture the same productivity measures, and they function under different productivity and training dynamics. Similarly, informal or less structured training procedures are frequently

associated with household-based work, making it challenging to evaluate productivity using conventional criteria. Finally, diplomatic and international institutions fall under the category of extraterritorial organizations, whose productivity is measured using other approaches that might not be the same as those employed in business-oriented industries.

Any analysis of how training affects productivity in these areas would be entirely hypothetical given these limitations. In many instances, the lack of statistical significance emphasizes the limits of the available data in capturing potential effects rather than necessarily indicating the absence of a relationship.

Sectors with no statistically significant correlation.

Training and productivity have been shown to be partially or strongly correlated in some industries, although there is no statistically significant association between the investigated regressors and productivity in several other businesses. This implies that variables other than training might be more important in determining productivity levels in these industries.

There may be several reasons for this lack of correlation, including the industry's structural features, legal restrictions, or the nature of the labor itself, which may not lend itself to quantifiable productivity gains through training. Sectors that primarily depend on capital investment rather than human labor, for example, may not respond well to training programs. Similarly, traditional regression models could not show productivity benefits right away in businesses with long-term skill development procedures or highly specialized knowledge requirements.

These industries are briefly discussed in the sections that follow, with an emphasis on potential externalities for the lack of link. The appendix contains the complete Stata code and robust regression results for completeness (from *Code.14* to *Code.21*).

There are no statistically significant relationships between training variables and productivity in the regression study for the *supply sectors of gas, steam, electricity, and air conditioning* (ATECO 35). The p-values are constantly over 0.1, indicating that none of the coefficients are significant. With a probability (Prob > F) of 0.3380, the overall F-test demonstrates that there is no association and that the model has limited ability to explain variance in productivity. The sector's high capital requirements could be one reason for this lack of correlation. Rather than worker training, infrastructure and technology improvements frequently have a greater impact on productivity in these operations.

The corresponding Stata code and regression output can be found in the appendix as *Code.14*.

Additionally, regression results show no statistically significant relationships between training-related variables and productivity in the *water supply and waste management* sectors (ATECO 36-

39). The model appears to be insufficient in explaining fluctuations in productivity, as indicated by the Prob > F value of 0.4966. The only coefficient that comes close to a little relevance (p = 0.110) is the one for the percentage of trained workers (`trained_quota100`), but it falls short of the traditional significance thresholds (p < 0.05). The industry's heavy reliance on physical infrastructure and stringent environmental standards may be one factor contributing to this lack of link. Investments in waste treatment and purification facilities probably have a greater impact on operational efficiency than employee training.
The corresponding Stata code and regression output can be found in the appendix as *Code.15*.

Regression analysis reveals no statistically significant correlation between productivity and training-related variables in the *accommodation and food service* industries (ATECO 55-56). The model has little explanatory power regarding productivity in this sector, as evidenced by the highest Prob>F value of 0.7375 among those examined thus far. The absence of statistical significance is confirmed by the p-values for each coefficient being more than 0.3. The nature of the industry, which is marked by significant labour turnover and extreme seasonality, may be the cause of this lack of linkage. Enhancing employee abilities is essential for providing high-quality services, but using the metrics used in this analysis, it could not result in quantifiable increases in productivity.
The corresponding Stata code and regression output can be found in the appendix as *Code.16*.

There are no statistically significant correlations between training-related factors and productivity, according to the *information and communication services* sectors' (ATECO 58-63) regression analysis. The model has virtually minimal explanatory power, as indicated by the Prob > F value of 0.9075. Furthermore, the p-values of all the coefficients are substantially higher than the traditional significance thresholds, indicating that training and the other factors considered have no discernible impact on productivity in this industry. The industry's dependence on innovation and technology developments rather than conventional personnel training could be one reason. Although employee skills are important, they are often acquired through informal on-the-job learning or self-taught means, which reduces the quantifiable value of formal training.
The corresponding Stata code and regression output can be found in the appendix as *Code.17*.

The regression results for the *insurance and finance* industries (ATECO 64-66) indicate a limited association between training-related variables and productivity. The Prob > F value of 0.0667 for the entire model is marginally close to significance but still over the conventional 0.05 cutoff. The sole independent variable that achieves significance is `vB15_2` (p = 0.044); however, the impact magnitude is modest. There is no statistically significant effect of the remaining variables on production. The

fact that financial and insurance operations depend more on risk assessment, regulatory compliance, and sophisticated data analytics than just employee training could be one explanation.
The corresponding Stata code and regression output can be found in the appendix as *Code.18*.

Compared to the other areas studied up to this point, the *real estate* industry (ATECO 68) shows some discrepancies. The regression model shows a Prob > F value equal 0.0378 which, in context, is moderately effective. Still, none of the individual coefficients are statistically significant which indicates that no definable productivity influencing variable correlates with training. Themsmall sample size of 14 observations is likely contributing to the fragility of the results.
Perhaps one suggestion to consider is that external market factors such as property demand, interest rates, and prevailing macroeconomic conditions drives productivity in real estate. While training employees may improve customer service and negotiation skills, it is highly improbable that it could directly and quantifiably affect productivity for an entire industry.
The corresponding Stata code and regression output can be found in the appendix as *Code.19*.

Training-related characteristics and productivity do not significantly correlate also for *professional, scientific, and technical* activities (ATECO 69-75). With a Prob > F = 0.3653, the model's overall explanatory power is poor, suggesting that the independent variables do not meaningfully explain productivity fluctuations taken together. With a p-value of 0.061, `vB11` is close to significance, but it still falls short of accepted limits.
One explanation could be that this industry contains highly specialized occupations where knowledge, experience, and intellectual capital (rather than formal education) have an impact on productivity. Short-term training has less of an influence because many experts in this industry, including consultants, engineers, and researchers, develop their talents via years of practice and further schooling. Furthermore, because output in this sector may be qualitative rather than quantitative, it is frequently challenging to gauge productivity using conventional economic metrics.
The corresponding Stata code and regression output can be found in the appendix as *Code.20*.

There are no discernible relationships between training-related variables and productivity, according to the *education* sector (ATECO 85), too. The model appears to have almost no explanatory power, as indicated by the Prob > F value of 0.9281. None of the variables included have a significant impact on productivity in this industry, as indicated by the p-values of all the coefficients being much over 0.05. The difficulty of measuring educational productivity using conventional economic measures may be one factor contributing to this outcome. Rather than only training or personnel costs, a variety of factors, such as curriculum quality, student involvement, and institutional resources, affect

teaching effectiveness and student outcomes. Furthermore, rather than through formal training programs, educators frequently acquire abilities through experience and long-term practice, which reduces the quantifiable impact of short-term training efforts.
The corresponding Stata code and regression output can be found in the appendix as *Code.21*.

Another time, the regression results show no significant correlations between production and training for the *other service activities* sector (ATECO 94-96). The model appears to be insufficient in explaining fluctuations in productivity, as indicated by the Prob > F value of 0.8151. With p-values significantly higher than typical thresholds, none of the coefficients are statistically significant.
The diverse character of this industry, which encompasses a range of service-based businesses like associations, personal care, and repair services, may be the cause of this outcome. Formal staff training is unlikely to have as much of an impact on productivity drivers in these industries as client demand, business strategies, and operational efficiency. Furthermore, a lot of service-oriented positions need for practical experience and the ability to engage with customers, which conventional training metrics might not adequately measure.
The corresponding Stata code and regression output can be found in the appendix as *Code.22*.

Analysis of the relationship between productivity and graduates in the workforce.
Following the assessment of training factors and their impact on productivity, the next endeavour is to evaluate whether a company's productivity is dependent on the percentage of its workforce that has completed tertiary education. The essential argument in this case would be how the level of education and skills attained enhances the productivity of the employees in the company.
The analysis concentrates on two aspects: the first deals with productivity and the proportion of employees with master's degrees (`vB3_1`) as a ratio of total staff, while the second aspect examines productivity and the percentage of employees with bachelor's and master's degrees (`vB3_1 + vB3_2`) in relation to total employees.
This study is very useful in the context of Becker's Human Capital Theory, which postulates that spending on education and training ameliorates human capital and, therefore, increases productivity of the employees. In Becker's terms, more educated workers like holders of university qualifications are assumed to be more productive, creative, and flexible which leads to better results at the firms. Hence, it is important to analyse whether more educated employees translate to greater productivity to appreciate the role of human capital on firms' performance.
To perform this analysis, the new employee variable `graduates_quota100` represents the percentage of employees with a bachelor's degree or higher. It is computed by adding the number of employees

with masters (vB3_1) and bachelor's degrees (vB3_2), dividing by the total number of employees (_v1), and multiplying by one hundred. The Stata code to compute that variable is provided below:

```
gen graduates_quota100 = ((vB3_1 + vB3_2) / _v1) * 100
```

In addition, a second variable, bachelors_quota100, will be created to define the percentage of an organization's employees with a master's degree out of the total employees. This will be done by taking the number of employees with master's degrees (vB3_1) over total employees (_v1) and multiplying the outcome by 100. The code in Stata that does this is:

```
gen masters_quota100 = (vB3_1 / _v1) * 100
```

After these variables are created, they will be incorporated alongside a regression analysis to check their relationship with productivity. This will help to test whether the suggest by the Becker theory: the more highly educated employees such as bachelor's or master's holders there are, the higher the productivity gets. The analysis aims to ascertain whether organizations with a greater share of highly educated employees are more productive, thereby confirming the hypothesis that human capital is vital to organizational performance.
The following *Table 13* presents the results of the command

```
reg productivity_hours bachelors_quota100 graduates_quota100
```

useful to ask Stata to do the regression with productivity_hours as dependent variable and masters_quota100 and graduates_quota100 as independent ones.

Source	SS	df	MS	Number of obs	=	13,171
				F(2, 13168)	=	2.06
Model	2.1024e+11	2	1.0512e+11	Prob > F	=	0.1269
Residual	6.7039e+14	13,168	5.0911e+10	R-squared	=	0.0003
				Adj R-squared	=	0.0002
Total	6.7060e+14	13,170	5.0919e+10	Root MSE	=	2.3e+05

productivity_hours	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
masters_quota100	257.004	185.359	1.39	0.166	-106.3264	620.3345
graduates_quota100	-38.53011	145.726	-0.26	0.791	-324.1742	247.114
_cons	3737.886	2299.836	1.63	0.104	-770.1242	8245.896

Table 13: regression output for productivity and graduates

The results from this regression analysis show some problems with the explanation. The R-squared value (0.0003) is particularly low which means the independent factors fail to explain the fluctuation in hourly productivity. That means there are other unsaid reasons that have much more impact if the

productivity levels. The accompanying F-statistic is also not significant (p = 0.1269) which means there is no clear indication that the model is helpful in explaining the observed data. The high Root MSE of 2.3e+05 also implies a high degree of scatter among the residuals which further puts the credibility of the model into question.

Looking at the remaining coefficients, the fraction of employees with bachelor's degrees relative to the total number of employees has a value of 257.004. This suggests productivity tends to increase as the portion of at least a bachelor's degree holders increases. However, the effect is not statistically significant with a p-value of 0.166 which is greater than standard cutoff meaning the effect observed is simply random fluctuation. The co-efficient of the fraction of employees with a master's degree relative to the total number of employees also has a value of – 38.53011. While this may imply that a higher number of master's graduates corresponds with lower production, the extraordinarily high p-value (0.791) indicates that this conclusion is far from significant and should not be taken seriously.

Moreover, the intercept is estimated to be 3737.886, corresponding to the expected level of productivity when both independent variables equal zero. In addition, this value is not statistically significant (p = 0.104), which points back to the fundamental problem that none of the estimated coefficients are telling us much, if anything at all, about the relationship between the share of graduates and firm productivity.

These results indicate that the model has a number of econometric problems. The very low explanatory power suggests that some of the most important determinants of productivity are left out of this model. There is a great deal of heteroskedasticity that most probably biases the estimates, and this could be verified by formal testing. Also, the large Root MSE makes it likely that some of the data is contaminated with outliers, which would be expected to bias the results. In addition, weak relevance of some variables' estimates explains not only how potential but also why so few statistical characteristics are significant. Such elements as working experience, industry type, firm size, and investment in innovation and training are highly likely to determine productivity but are excluded from this study.

Given these constraints, a more robust regression approach is required to produce reliable results. A heteroskedasticity test and an assessment of the influence of outliers would both be useful milestones in model refinement. However, the goal of this thesis is not to do a thorough econometric analysis, but rather to provide a beginning exploration of the relationship between productivity and the proportion of graduates in the workforce. The question is posed as a point of interest rather than the primary focus of the study.

For completeness and curiosity, a robust regression was requested from Stata using

```
rreg productivity_hours bachelors_quota100 graduates_quota100
```

to see if the existence of outliers or heteroskedasticity was influencing the results. This approach aims to reduce the influence of extreme values and any breaches of traditional regression principles, as already stated. However, as previously noted, the major goal of this thesis is not to do thorough econometric research of this relationship, but rather to present an early exploratory analysis of the relationship between productivity and the proportion of graduates in the labour force. Therefore, the data, shown in *Figure 19*, should be considered as a preliminary suggestion rather than definitive evidence.

```
2 . rreg productivity_hours masters_quota100 graduates_quota100

    Huber iteration 1:  maximum difference in weights = .99994178
    Huber iteration 2:  maximum difference in weights = .98514142
    Huber iteration 3:  maximum difference in weights = .86748903
    Huber iteration 4:  maximum difference in weights = .6571912
    Huber iteration 5:  maximum difference in weights = .45405602
    Huber iteration 6:  maximum difference in weights = .26316322
    Huber iteration 7:  maximum difference in weights = .09308321
    Huber iteration 8:  maximum difference in weights = .02500808
  Biweight iteration 9:  maximum difference in weights = .29418045
  Biweight iteration 10:  maximum difference in weights = .097817
  Biweight iteration 11:  maximum difference in weights = .02493686
  Biweight iteration 12:  maximum difference in weights = .00550911

Robust regression                               Number of obs     =     13,171
                                                F(  2,     13168) =       8.38
                                                Prob > F          =     0.0002

------------------------------------------------------------------------------
productivity_hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
  masters_quota100 |   .2195509   .069568     3.16   0.002     .0831875    .3559143
graduates_quota100 |  -.0575136   .0546932   -1.05   0.293    -.1647201     .049693
             _cons |   104.1847   .863163   120.70   0.000     102.4928    105.8766
------------------------------------------------------------------------------
```

Figure 19: robust regression output for graduates' quotas

The findings of the robust regression show several noticeable deviations from the Ordinary Least Squares (OLS) regression. Specifically, the coefficient for the proportion of employees with a bachelor's degree in the workforce (bachelors_quota100) is now statistically significant at the 1% level, indicating a positive effect on productivity per hour worked. Specifically, the coefficient of 0.2196 indicates that a one-point rise in the share of bachelor's degree holders inside a corporation is connected with a 0.22 increase in productivity per hour, holding other variables constant. In contrast, the coefficient for the proportion of employees with a graduate degree (graduates_quota100) remains statistically insignificant, meaning that there is no clear evidence that a higher share of master's degree holders has a consistent effect on productivity.

The robust regression produces a statistically significant result for bachelors_quota100, implying that the initial OLS findings were influenced by outliers or heteroskedasticity. The Huber and Biweight iterations show incremental convergence in weight modifications, which reinforces the estimate process's resilience. This demonstrates how typical regression algorithms can fail to capture certain correlations due to outliers or deviations from homoscedasticity.

Several variables could explain why the variable indicating graduate degree holders is not significant. One hypothesis is that organizations with a higher share of people with advanced degrees operate in industries where productivity is difficult to assess using the existing information. Alternatively, a non-linear relationship between education level and productivity may exist, implying that beyond a certain point, more formal education does not always translate into increased production. Another possible explanation is that master's degree holders are more likely to be employed in roles with a less direct impact on immediate business productivity, such as research, management, or administration.

It would be useful to broaden the analysis by looking at the association between education levels and productivity across different sectors, using the ATECO categorization as previously used. A sectoral split may reveal varied effects, as education's influence is expected to vary based on industry characteristics. However, such an analysis is outside the scope of this thesis. Instead, this study provides an early glimpse into the relationship between the proportion of university-educated staff and production. Further research would be required to investigate sector-specific patterns and discover potential causes underlying the observed trends.

6. CONCLUSIONS AND IMPLICATIONS

The relationship between on-the-job training and productivity was investigated using data from INAPP's 2018 RIL questionnaire. While INAPP offered data from several polls conducted throughout time, sensitive company-level data was unavailable, prohibiting the development of a comprehensive panel for a more detailed longitudinal analysis. As a result, the study was confined to examining data from the 2018 survey alone. This limitation, while necessary, created significant challenges to the study, as it was anticipated from the start that establishing a direct relationship between on-the-job training and productivity within a single time would be difficult.

For that reason, this study used a cross-sectional technique, rather than longitudinal data, to investigate the association between training and productivity at a certain moment. This decision, while limiting the ability to detect temporal changes, allowed for a precise snapshot of the relationship between production and training inside the data set. The model considered four essential criteria for on-the-job training: total training cost, number of employees trained, percentage of employees trained, and training funding per employee.

Becker's human capital theory serves as the basis for understanding this problem. According to Becker, the advantages of on-the-job training frequently emerge gradually rather than immediately. His research concludes: *"Training might lower current receipts and raise current expenditures, yet firms could profitably provide this training if future receipts were sufficiently raised, or future expenditures sufficiently lowered."* This study underscores the fact that the advantages of on-the-job training are linked to future periods and hence cannot be successfully measured in a single timeframe. As a result, the assumption of a clear and direct correlation between training and production in the short term seemed fair from the outset.

Throughout the thesis, the study by Dearden et al. (2006) was examined and elaborated upon, reinforcing the view that the relationship between training investments and productivity is complex and may take time to emerge. Dearden et al. (2006) go beyond the standard way of utilizing wages as a reliable indication of productivity by directly measuring the impact of work-related training on it. Dearden et al. also discuss the significance of externalities in training, which are often overlooked in individual-level assessments. They remark that the returns to training may be higher at the industry level, as the overall impact of training may include positive spillover effects that benefit other businesses and sectors within the industry. This externality impact is crucial because it means that the full benefits of training extend beyond the individual worker or business and influence the whole economic environment. Their research also revealed that training externalities may account for a

significant portion of the productivity improvements reported at the industry level, and at first glance it is what drives productivity for the sectors discussed in *Sectors with no statistically significant correlation*.

The first regression analysis found that, at standard significance levels, none of the main explanatory variables (training expense, number of trained employees, percentage of trained employees, or training expenditure per employee) had a statistically significant impact on productivity. Given these findings, additional tests were required to determine the appropriateness of utilizing robust regression instead of the usual Ordinary Least Squares (OLS) technique. The existence of heteroscedasticity and influential outliers in the dataset rendered the OLS estimates inaccurate, hence the robust regression technique was used to address these difficulties. Robust regression produced more credible estimates by controlling for heteroscedasticity and lowering the influence of extreme observations, resulting in a more accurate portrayal of the link between training and productivity.
The robust regression analysis yielded considerable findings. To begin, it was discovered that the number of trained employees has a statistically significant positive association with productivity, implying that increasing the absolute number of trained workers results in increased production. This contrasted with the OLS findings, which revealed no discernible link. On the other hand, the percentage of trained personnel showed a negative association with production, implying that having more skilled employees may impair output. This statistically significant conclusion implies that, while educating a larger workforce can be useful to some extent, over-investment in training may not result in equal productivity increases. Furthermore, the research supported the OLS finding that training expenditures per employee had no significant influence on production.

Additionally, sector-specific analysis found that the relationship between training and production varied significantly by industry. In the manufacturing industry, the quantity of skilled workers is a primary driver of productivity, whereas training expenses and per-employee costs have no discernible effect. In contrast, in the construction sector, training expenditures have a negative impact on productivity, presumably due to inefficient allocation or operational disturbances, even though the absolute number of trained people contributes favourably. Training expenditures and the proportion of trained staff are positively related to productivity in wholesale and retail trade, as well as motor vehicle and motorcycle repair, highlighting the importance of a professional workforce in customer-centric contexts. The transportation and storage industry indicates that training quality and targeted investments are more significant than just expanding the quantity of skilled people. While more trained individuals increase production in the arts and entertainment industry, bigger training

expenditures appear to be counterproductive, showing a dependence on specialized skills rather than formal training programs.

Instead, the statistics show that in some other industries the link between training and productivity is weak or non-existent. For example, in mining and quarrying, there is no substantial association; rather than training, business size drives productivity, most likely due to the sector's capital-intensive character. Similarly, in the rental, travel agency, and business support services industries, labour expenses are the only significant predictors of productivity, indicating that operational efficiencies and customer service outweigh the effects of formal training. In the healthcare and social work sectors, an unanticipated negative association is noted, probably because training is provided in response to regulatory changes or talent shortages, temporarily lowering output. Furthermore, regression analyses in some other sectors[31] show that training has little to no significant impact on productivity. In many circumstances, structural characteristics, legal frameworks, and external factors such as technical investments, customer demand, or market conditions appear to have a greater influence than on-the-job training. The absence of statistically significant connections in certain businesses is due to data limitations and the dynamics of each sector, rather than a complete lack of influence.

Furthermore, an exploratory analysis was carried out to determine whether the percentage of university-educated staff effects productivity. When a robust regression technique was used, the fraction of employees with a bachelor's degree was shown to be statistically significant and positively connected with productivity, demonstrating that having a bachelor's degree is linked to higher hourly production. In contrast, the fraction of employees with a master's degree remained statistically insignificant, assuming that beyond a certain educational threshold, more formal education does not always translate into increased productivity. These data imply that, while the level of formal education may influence productivity, the effects are likely subtle and dependent on other factors not represented by this model. Clearly, these statements require a dedicated and in-depth study.

In conclusion, while the study demonstrates a clear link between on-the-job training and productivity, it is important to note that the analysis does not establish causation due to the constraints of cross-sectional data. The methods employed, notably the use of robust regression, improved the findings' credibility by addressing potential flaws with the data. Future studies should include longitudinal data

[31] such as energy supply, water supply, waste management, accommodation and food services, information and communication services, finance and insurance, real estate, professional, scientific and technical activities, education, and other service activities

to better understand the long-term effects of training on productivity and business performance. Furthermore, as proposed by Dearden et al. (2006), investigating potential externalities and doing more in-depth industry-level evaluations could provide additional insights into training's overall economic impact.

References

Ackerberg, D., Caves, K., & Frazer, G. (2006). Structural identification of production functions.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University press.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, *58*(2), 277-297.

Becker, G. S. (1964). Human capital theory.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics*, *87*(1), 115-143.

Bratti, M., Conti, M., & Sulis, G. (2018). *Employment protection, temporary contracts and firm-provided training: Evidence from Italy* (No. 2018/2). JRC Working Papers in Economics and Finance.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata* (Vol. 2). College Station, TX: Stata press.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (No. 53). Cambridge University press.

Chesher, A. (1979). Testing the law of proportionate effect. *The Journal of industrial economics*, *27*(4), 403-411.

Costa, S., De Santis, S., & Monducci, R. (2022). Reacting to the COVID-19 crisis: State, strategies and perspectives of Italian firms. *Rivista di statistica ufficiale/Review of official statistics*, *1*, 73-107.

De Grip, A., & Sauermann, J. (2013). The effect of training on productivity: The transfer of on-the-job training from the perspective of economics. *Educational Research Review*, *8*, 28-36.

Dearden, L., Reed, H., & Van Reenen, J. (2006). The impact of training on productivity and wages: Evidence from British panel data. *Oxford bulletin of economics and statistics*, *68*(4), 397-421.

Ferri, V., Ricci, A., & Sacchi, S. (2018). Demografia imprenditoriale e tessuto produttivo in Italia.

Michael, R. T. (1972). *The effect of education on efficiency in consumption* (No. mich72-1). National Bureau of Economic Research.

Penrose, E. T. (2009). *The Theory of the Growth of the Firm*. Oxford University press.

Psacharopoulos, G., & Patrinos*, H. A. (2004). Returns to investment in education: a further update. *Education economics*, *12*(2), 111-134.

Schultz, T. W. (1961). Investment in human capital. *The American economic review*, *51*(1), 1-17

Sitography

https://www.istat.it/wp-content/uploads/2022/03/volume_integrale_ATECO2007.pdf

https://www.bancaditalia.it/pubblicazioni/relazione-annuale/2018/sintesi/index.html

https://www.panorama.it/economia/la-classifica-delle-imprese-italiane-per-fatturato

Appendix

Table A.1

VARIABILE	missing	obs	mean	st.dev.	min	max
CASENUM	0	30023	15012	8666,894	1	30023
wcal	0	30023	52,6185	229,7521	3,22E-38	5835,819
VA4	0	30023	/	/	0	0
VA5	0	30023	/	/	0	0
Ripartizione	0	30023	/	/	0	0
ATECO2007	4	30019	/	/	0	0
vA9	0	30023	2,133598	26,14586	0	3773
vA11	0	30023	/	/	0	0
_v1	0	30023	57,93282	327,7737	0	36169
_v2	1	30022	20,87076	162,2365	0	17042
_v3	4	30019	0,990773	8,174254	0	774
_v4	4	30019	0,152936	1,691671	0	156
_v5	4	30019	3,543056	111,4076	0	18377
_v6	4	30019	1,060129	41,82975	0	6944
_v7	0	30023	23,92336	184,9709	0	17018
_v8	0	30023	11,67052	104,0379	0	9983
_v9	4	30019	29,41893	142,2231	0	9749
_v10	3	30020	8,020553	79,6114	0	8291
vB2	4	30019	63,48199	1519,099	-2	256541
vB3_VERIFICA	8670	21353	/	/	0	0
vB3_1	8670	21353	8,516977	107,9717	0	13273
vB3_2	8670	21353	3,505315	37,87951	0	2731
vB3_3	8670	21353	24,62446	167,007	0	18956
vB3_4	8670	21353	18,27219	113,1354	0	9592
vB3B_1	24829	5194	8,938005	16,3732	0	100
vB3B_2	24829	5194	5,846746	14,17864	0	100
vB3B_3	24829	5194	49,25683	29,05703	0	100
vB3B_4	24829	5194	35,96034	31,50194	0	100
vB4_VERIFICA	7214	22809	/	/	0	0
vB4_1	7214	22809	2,80933	18,80264	0	900
vB4_2	7215	22808	11,21672	54,52702	0	2834
vB4_3	7214	22809	27,27792	168,4396	0	17542
vB4_4	7214	22809	15,57039	120,1056	0	14208
vB4_5	7214	22809	2,782104	21,1131	-1	1634
vB4BIS_1	26285	3738	6,222579	11,44411	0	100
vB4BIS_2	26285	3738	23,55752	21,628	0	100
vB4BIS_3	26285	3738	44,93365	24,38242	0	100
vB4BIS_4	26285	3738	21,86196	20,68169	0	100
vB4BIS_5	26285	3738	3,44168	8,277907	0	100
_v11	3477	26546	56,70139	324,2824	0	35301
_v12	3476	26547	20,15395	157,3992	0	16565
_v13	3476	26547	6,577805	36,7991	0	1616
_v14	3476	26547	2,557276	20,23992	0	1259
_v15	3476	26547	1,390477	14,11712	0	927
_v16	3476	26547	0,516819	7,511023	0	599
_v17	3476	26547	0,817079	41,33764	0	5725
_v18	3476	26547	0,411723	25,69717	0	3911
_v19	3474	26549	11,02787	110,9454	0	8890
_v20	3474	26549	7,913405	91,22488	0	8125
_v21	3474	26549	9,232928	97,72257	0	7933
_v22	3474	26549	6,831293	83,05161	0	7329
_v23	3474	26549	1,794945	20,32001	0	1242
_v24	3474	26549	1,08215	13,12836	0	796
_v25	1	30022	1,115815	13,44433	0	1850
_v26	1	30022	0,329858	7,94015	0	1100
_v27	1	30022	0,565918	7,618823	0	601
_v28	1	30022	0,16418	3,288758	0	280
_v29	1	30022	1,659516	32,21505	0	4253
_v30	1	30022	0,203584	6,544462	0	799
_v31	1	30022	0,034841	0,690629	0	100
_v32	1	30022	0,014023	0,164974	0	10
_v33	1	30022	0,478083	4,722868	0	394
_v34	1	30022	0,221105	2,64749	0	195
_v35	0	30023	0,21267	5,512985	0	400
_v36	0	30023	0,081637	3,263893	-1	380
vB10	0	30023	/	/	0	0
vB11	14189	15834	60,87129	389,4031	0	37843
vB11A	15794	14229	/	/	0	0
_v37	18890	11133	52,49735	405,9728	0	37006
_v38	18890	11133	17,15989	183,0018	0	17152
_v39	18890	11133	8,973143	301,8228	0	30123

_v40	18890	11133	2,163388	33,89904	0	1946
_v41	18890	11133	2,192221	15,84545	0	522
_v42	18890	11133	0,751819	8,158226	0	340
_v43	18890	11133	3,693075	27,00133	0	1250
_v44	18890	11133	1,482979	17,50536	-3	1125
vB12_M_1	14180	15843	/	/	0	0
vB12_M_2	14180	15843	/	/	0	0
vB12_M_3	14180	15843	/	/	0	0
vB12_M_4	14180	15843	/	/	0	0
vB12_M_5	14185	15838	/	/	0	0
vB13	14180	15843	/	/	0	0
vB14_M_1	24738	5285	0,695175	0,460333	0	1
vB14_M_2	24738	5285	0,271902	0,444939	0	1
vB14_M_3	24738	5285	0,108231	0,310672	0	1
vB15	18603	11420	1	0	1	1
vB15_2	18603	11420	77328,2	2241338	0	2E+08
vD1	17433	12590	/	/	0	0
vD2	11747	18276	/	/	0	0
vD3	24014	6009	/	/	0	0
vD4	20205	9818	/	/	0	0
vD5	26581	3442	/	/	0	0
vC1	0	30023	/	/	0	0
vC2	14454	15569	17,75856	104,6399	0	9234
vC3	14388	15635	/	/	0	0
vC4_M_1	26080	3943	/	/	0	0
vC4_M_2	26080	3943	/	/	0	0
vC4_M_3	26080	3943	/	/	0	0
vC4_M_4	26080	3943	/	/	0	0
vC4_M_5	26080	3943	/	/	0	0
vC4_M_6	26080	3943	/	/	0	0
vC4_M_7	26080	3943	/	/	0	0
vC4_M_8	26080	3943	/	/	0	0
vC4_M_9	26080	3943	/	/	0	0
vC4_M_10	26080	3943	/	/	0	0
vC4BIS	26080	3943	/	/	0	0
vC5	0	30023	/	/	0	0
vC6_1	14726	15297	16,82794	106,5416	0	9246
vC6_2	14726	15297	1,709159	10,94136	0	378
vC6_3	14726	15297	0,783683	3,67722	0	198
vC6_4	14726	15297	0,332353	26,84618	0	3304
vC6_5	14726	15297	8,641106	86,82373	0	8572
vC6_6	14726	15297	4,258286	17,74893	0	736
vC6_7	14726	15297	1,103354	14,06222	-1	713
vC7	0	30023	/	/	0	0
vC7BIS	25071	4952	5,452746	17,97017	0	500
vC8_M_1	25051	4972	/	/	0	0
vC8_M_2	25051	4972	/	/	0	0
vC8_M_3	25051	4972	/	/	0	0
vC8_M_4	25051	4972	/	/	0	0
vC8_M_5	25051	4972	/	/	0	0
vC8_M_6	25051	4972	/	/	0	0
vC8_M_7	25051	4972	/	/	0	0
vC8_M_8	25051	4972	/	/	0	0
vC9	0	30023	/	/	0	0
vC10_1	26407	3616	0,746128	4,682182	0	130
vC10_2	26407	3616	9,003319	47,68575	0	1958
vC10_3	26407	3616	0,488938	3,810412	-3	130
vC11_M_1	0	30023	/	/	0	0
vC11_M_2	0	30023	/	/	0	0
vC11_M_3	0	30023	/	/	0	0
vC11_M_4	0	30023	/	/	0	0
vC11_M_5	0	30023	/	/	0	0
vC12	0	30023	/	/	0	0
vC13_M_1	0	30023	/	/	0	0
vC13_M_2	0	30023	/	/	0	0
vC13_M_3	0	30023	/	/	0	0
vC14	0	30023	/	/	0	0
vC14A	29648	375	4,922667	8,756979	0	100
vC14B	29276	747	9,528782	21,35491	0	300
vC15	0	30023	/	/	0	0
vC15A	29753	270	4,292593	11,52419	0	100
vC15B	28905	1118	8,031306	18,04789	0	308
vC16	0	30023	/	/	0	0
vF1	0	30023	/	/	0	0
vF1BIS	28296	1727	2,193978	0,671195	0	11
vF2	3474	26549	/	/	0	0
vF3	8130	21893	1,072854	0,259897	1	2
vF3_2	0	30023	/	/	0	0

vF3B	0	30023	/	/	0	0
vF4	3471	26552	/	/	0	0
vF5	26764	3259	/	/	0	0
vF6_M_1	26764	3259	/	/	0	0
vF6_M_2	26764	3259	/	/	0	0
vF6_M_3	26764	3259	/	/	0	0
vF6_M_4	26764	3259	/	/	0	0
vF6_M_5	26764	3259	/	/	0	0
vF6_M_6	26764	3259	/	/	0	0
vF6_M_7	26764	3259	/	/	0	0
vF6_M_8	26764	3259	/	/	0	0
vF6_M_9	26764	3259	/	/	0	0
vF6_M_10	26764	3259	/	/	0	0
vF6_M_11	26764	3259	/	/	0	0
vF7_M_1	26764	3259	/	/	0	0
vF7_M_2	26764	3259	/	/	0	0
vF7_M_3	26764	3259	/	/	0	0
vF7_M_4	26764	3259	/	/	0	0
vF7_M_5	26764	3259	/	/	0	0
vF8	27555	2468	/	/	0	0
vF9	28101	1922	/	/	0	0
vF10	3474	26549	/	/	0	0
vF11	24575	5448	68,19035	449,292	0	28219
vF12	4826	25197	1541,285	233328,7	-3	37034000
vF13	3471	26552	/	/	0	0
vF14	29046	977	/	/	0	0
vL1	0	30023	/	/	0	0
vL2	0	30023	/	/	0	0
vL3	0	30023	/	/	0	0
vL4_1	0	30023	/	/	0	0
vL4_2	0	30023	/	/	0	0
vL4_3	0	30023	/	/	0	0
vL4_4	0	30023	/	/	0	0
vL4_5	0	30023	/	/	0	0
vL4_6	0	30023	/	/	0	0
vL4_7	0	30023	/	/	0	0
vL6	0	30023	/	/	0	0
vL7	22476	7547	38,03684	138,8487	-1	11412
vL8_1	321	29702	/	/	0	0
vL8_2	306	29717	/	/	0	0
vL9	0	30023	/	/	0	0
vL10	354	29669	/	/	0	0
vL11	22204	7819	22,9468	30,81404	0	100
vH1	444	29579	/	/	0	0
vH2	19039	10984	4426326	1,18E+08	0	9,3E+09
vH3	17442	12581	/	/	0	0
vH4_M_1	23601	6422	/	/	0	0
vH4_M_2	23601	6422	/	/	0	0
vH4_M_3	23601	6422	/	/	0	0
vH4_M_4	23601	6422	/	/	0	0
vH4_M_5	23601	6422	/	/	0	0
vH4_M_6	23601	6422	/	/	0	0
vH4_M_7	23601	6422	/	/	0	0
vH4_M_8	23601	6422	/	/	0	0
vH4BIS	23601	6422	/	/	0	0
vH5_M_1	17442	12581	/	/	0	0
vH5_M_2	17442	12581	/	/	0	0
vH5_M_3	17442	12581	/	/	0	0
vH5_M_4	17442	12581	/	/	0	0
vH5_M_5	17442	12581	/	/	0	0
vH5_M_6	17442	12581	/	/	0	0
vH6	0	30023	/	/	0	0
vH7	22763	7260	/	/	0	0
vH8	0	30023	/	/	0	0
vH9	24659	5364	/	/	0	0
vH10_M_1	24659	5364	/	/	0	0
vH10_M_2	24659	5364	/	/	0	0
vH10_M_3	24659	5364	/	/	0	0
vH10_M_4	24659	5364	/	/	0	0
vH10_M_5	24659	5364	/	/	0	0
vH10_M_6	24659	5364	/	/	0	0
vH11	0	30023	/	/	0	0
vH12	10904	19119	227229,6	5972200	-1	7,06E+08
vH13	4060	25963	4,19E+10	5,2E+12	-1000000	7,83E+14

Table A.2

diversification

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	17,630
25%	0	0	Sum of Wgt.	17,630
50%	0		Mean	.5117981
		Largest	Std. Dev.	.8541971
75%	1	4		
90%	2	4	Variance	.7296526
95%	2	4	Skewness	1.707221
99%	3	4	Kurtosis	5.290015

Table A.3

trained_quota100

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	14,159
25%	0	0	Sum of Wgt.	14,159
50%	26.08696		Mean	41.40993
		Largest	Std. Dev.	42.69956
75%	96.2963	100		
90%	100	100	Variance	1823.253
95%	100	100	Skewness	.3446672
99%	100	100	Kurtosis	1.370005

Table A.4

expense_per_empl

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	8,356
25%	23.80141	0	Sum of Wgt.	8,356
50%	109.432		Mean	892.2511
		Largest	Std. Dev.	14636.03
75%	255.4444	333333.3		
90%	537.5	500000	Variance	2.14e+08
95%	900	511111.1	Skewness	35.94334
99%	4524.167	811700	Kurtosis	1577.904

Table A.5

ATECO2007

Percentiles		Smallest		
1%	10	1		
5%	10	3		
10%	16	8	Obs	17,627
25%	28	8	Sum of Wgt.	17,627
50%	43		Mean	43.66693
		Largest	Std. Dev.	20.91894
75%	56	96		
90%	70	96	Variance	437.6022
95%	85	96	Skewness	.4097977
99%	96	96	Kurtosis	2.709307

Code.1

```
//summarize all the variables
estpost summarize
//Export data in excel to visualize better the data using colors and properties
export excel using "Datafile", sheet("dirtyData") firstrow(variables)
//drop all the negative values, that are invalid for the respective variable
drop if vB2<0 | vB4_5<0 | _v36<0 | _v44<0 | vC6_7<0 | vC10_3<0 | vF12<0 | vL7<0
| vH12<0 | vH13<0
//drop some useless variables
drop vB3_VERIFICA vB4_VERIFICA
drop vB4BIS_1 vB4BIS_2 vB4BIS_2 vB4BIS_3 vB4BIS_4 vB4BIS_5
drop _v25 _v26 _v27 _v28 _v29 _v30 _v31 _v32 _v33 _v34 _v35 _v36
drop vC4_M_* vC4BIS
drop vC10_1 vC10_2 vC10_3
drop vC11_M_*
drop vC12 vC13_M_* vC14 vC14* vC15 vC15* vC16
drop vF7_M_* vF13 vF14
drop vL8_*
drop vH4_M_*
drop vH10_M_*
drop vH11
//Analyse the variable _v1 (Total employees) and visualize it with an histogram
summarize _v1, detail
histogram _v1, title("Employees distribution") color(black) percent
//Analyse the variable _vH13 (Total sales for 2017) and visualize it with an
histogram
summarize vH13, detail
histogram vH13, title("Total sales distribution(2017)") color(black) percent
//Replace with missing the unacceptable values
replace vH13 = . if vH13>1e11
//Analyse the variable _vH13 (Total sales for 2017) and visualize it with an
histogram
summarize vH13, detail
histogram vH13, title("Total sales distribution (2017)") color(black) percent
//Replacing variables trackig the number of master's, bachelor's, diploma's and
compulsory education employees with 0 when _v1 is equal to 0
replace vB3_1=0 if _v1==0
replace vB3_2=0 if _v1==0
replace vB3_3=0 if _v1==0
replace vB3_4=0 if _v1==0
//Generate a variable to check the consistecy of the four variables before
w.r.t. the total number of employees
generate sum_vB3= vB3_1+vB3_2+vB3_3+vB3_4
drop if sum_vB3>_v1 | sum_vB3<_v1
//Knowing details about the composition of workforce in term of education
```

```
sum vB3_1, d
sum vB3_2, d
sum vB3_3, d
sum vB3_4, d
//Counting how much firms do not perform OJT
count if missing(vB11)
//Tabulate vB10, also without labels to know how replies were registered
tabulate vB10
tabulate vB10, nolabel
//Replace with 0 the values of vB11 for which vB10 has the answer 'No'
replace vB11 = 0 if vB10 == 2
//Replace with 0 the values of vB11 when the trained employees is 0 (_v1=0)
replace vB11 = 0 if _v1 == 0
//count and drop when the number of trained employees is high than total ones
count if vB11 > _v1
drop if vB11 > _v1
//Summarize vB11 with detail, and show an histogram graph
summarize vB11, detail
histogram vB11, title("Trained employees distribution (2017)") color(black)
percent
//Replacing the cost of training equal to 0 when missing
replace vB15_2 = 0 if vB10 == 1 & missing(vB15_2)
//Summarize vB15_2 with detail, and show an histogram graph
summarize vB15_2, detail
histogram vB15_2, title("Training expense distribution (2017)") color(black)
percent
//Cleaning if the worked hours per employee is higher than the maximum threshold
allowed
drop if vH12 / _v1 > 2496 & _v1>0
//Generate the productivity variable
generate productivity_hours = vH13/vH12
//Do an histogram of productivity and summarize it with detail
histogram productivity_hours, title("Productivity distribution (2017)")
color(black percent
summarize productivity_hours, detail
//Generate the diversification index and summarize it
gen diversification = (vB12_M_1 == 1) + (vB12_M_3 == 1) + (vB12_M_4 == 1) +
(vB12_M_5 == 1)
sum diversification, detail
//Generate the percentage of trained employees and summarize it
gen trained_quota100 = (vB11 / _v1) * 100
sum trained_quota100, detail
//Generate the expense per employee variable and summarize it
gen expense_per_empl = vB15_2 / _v1
sum expense_per_empl, detail
//Destring the ATECO variable and summarize it
destring ATECO2007, replace
sum ATECO2007, detail
//Regress for at least one type of OJT done
reg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1
//Scatterplot: productivity vs training expense
scatter productivity_hours vB15_2 if diversification>=1, title("Productivity vs.
Training expense") msize(small) mcolor(black)
//Scatterplot: productivity vs trained employees
scatter productivity_hours vB11 if diversification>=1, title("Productivity vs.
Trained employees") msize(small) mcolor(black)
//Scatterplot: productivity vs trained percentage
scatter productivity_hours trained_quota100 if diversification>=1,
title("Productivity vs. Trained percentage") msize(small) mcolor(black)
//Scatterplot: productivity vs training expense per employee
scatter productivity_hours expense_per_empl if diversification>=1,
```

```
title("Productivity vs. Training expense per employee") msize(small)
mcolor(black)
//Do the heteroscedasticity test
estat hettest
//Do the Cook's distance test for outliers
predict cooksd, cooksd
list cooksd if cooksd > 4/(_N) & cooksd != .
//Do the robust regression
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1
//Do the scatterplot for residuals and fitted values
predict fitted_values, xb
predict residuals, residuals
scatter residuals fitted_values, title("Fitted values vs. Residuals)
mcolor(black)
//Do the robust regression for the manufacturing sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=10 & ATECO2007<=33
//Do the robust regression for the construction sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 > & ATECO2007>=41 & ATECO2007<=43
//Do the robust regression for the retail sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 > & ATECO2007>=45 & ATECO2007<=47 ,
//Do the regression for the transportation and warehousing sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 > & ATECO2007>=49 & ATECO2007<=53
//Do the robust regression for the artistic, sports and entertainment activities
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=90 & ATECO2007<=93
//Do the regression for the mining and quarrying sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=5 & ATECO2007<=9
//Do the robust regression for the rental, travel agencies, and business support
services
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=77 & ATECO2007<=82
//Do the robust regression for the health and social work activities
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=86 & ATECO2007<=88
//Do the robust regression for the agriculture, forestry and fishing sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=01 & ATECO2007<=03
//Do the robsut regression for the public administration and defense, compulsory
social security sector
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007==84
//Do the robust regression for the households as employers of domestic personnel
sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=97 & ATECO2007<=98
//Do the robust regression for the extraterritorial organizations and bodies
sector
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007==99
//Do the robust regression for the supply sectors of gas, steam, electricity,
and air conditioning
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007==35
//Do the robust regression for the water supply and waste management sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=36 & ATECO2007<=39
```

```
//Do the robust regression for the accommodation and food service sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=55 & ATECO2007==56
//Do the robust regression for the information and communication services
sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=58 & ATECO2007<=63
//Do the robust regression for the insurance and finance sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=64 & ATECO2007<=66
//Do the robust regression for the real estate sector
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007==68
//Do the robsut regression for the professional, scientific, and technical
sectors
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=69 & ATECO2007<=75
//Do the robust regression for the education sector
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007==85
//Do the robust regression for the other service activities
rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if
diversification>=1 & ATECO2007>=94 & ATECO2007<=96
//Generate the variable tracking the number of bachelor's or higher degree in
the firm
gen graduates_quota100 = ((vB3_1 + vB3_2) / _v1) * 100
//Generate the variable tracking the number of master's degree in the firm
gen masters_quota100 = (vB3_1 / _v1) * 100
//Do the simple regression
reg productivity_hours bachelors_quota100 graduates_quota100
//Do the robust regression
rreg productivity_hours bachelors_quota100 graduates_quota100
```

Code.2

```
1 . tab vB10
```

Nel corso del 2017 sono state organizzate iniziative di formazione per i dipende	Freq.	Percent	Cum.
Si	12,607	50.81	50.81
No	12,205	49.19	100.00
Total	24,812	100.00	

```
2 . tab vB10, nolabel
```

Nel corso del 2017 sono state organizzate iniziative di formazione per i dipende	Freq.	Percent	Cum.
1	12,607	50.81	50.81
2	12,205	49.19	100.00
Total	24,812	100.00	

Code.3

```
7 . replace vB11=0 if _v1==0
  (80 real changes made)

8 . count if vB11 > _v1
    370

9 . drop if vB11 > _v1
  (370 observations deleted)
```

Code.4

```
2 . sum vB15_2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
vB15_2	9,131	77232.54	2443014	0	2.00e+08

Code.5

```
7 . generate productivity_hours = vH13/vH12
  (10,252 missing values generated)
```

Code.6

```
8 . drop if vH12/_v1>2496 & _v1>0
  (6,812 observations deleted)
```

Code.7

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1
  > & ATECO2007>=5 & ATECO2007<=9

   Huber iteration 1:  maximum difference in weights = .92654826
   Huber iteration 2:  maximum difference in weights = .73388101
   Huber iteration 3:  maximum difference in weights = .27249921
   Huber iteration 4:  maximum difference in weights = .09661779
   Huber iteration 5:  maximum difference in weights = .00304447
Biweight iteration 6:  maximum difference in weights = .19941894
Biweight iteration 7:  maximum difference in weights = .06990478
Biweight iteration 8:  maximum difference in weights = .02445745
Biweight iteration 9:  maximum difference in weights = .00160257

Robust regression                               Number of obs     =         36
                                                F(  4,        31) =       5.69
                                                Prob > F          =     0.0015
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0021514	.0014144	-1.52	0.138	-.005036	.0007332
vB11	2.42276	.6980931	3.47	0.002	.9989902	3.846531
trained_quota100	-.42899	.4049181	-1.06	0.298	-1.254826	.3968458
expense_per_empl	.0763991	.0760975	1.00	0.323	-.0788028	.2316011
_cons	104.6807	24.52984	4.27	0.000	54.65171	154.7096

Code.8

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7>=77 & ATECO2007<=82

   Huber iteration 1:  maximum difference in weights = .98517268
   Huber iteration 2:  maximum difference in weights = .7805658
   Huber iteration 3:  maximum difference in weights = .43306726
   Huber iteration 4:  maximum difference in weights = .2472065
   Huber iteration 5:  maximum difference in weights = .43943815
   Huber iteration 6:  maximum difference in weights = .16418455
   Huber iteration 7:  maximum difference in weights = .07107013
   Huber iteration 8:  maximum difference in weights = .02661603
Biweight iteration 9:  maximum difference in weights = .29202905
Biweight iteration 10:  maximum difference in weights = .24107876
Biweight iteration 11:  maximum difference in weights = .08742611
Biweight iteration 12:  maximum difference in weights = .03896648
Biweight iteration 13:  maximum difference in weights = .01632134
Biweight iteration 14:  maximum difference in weights = .00667391

Robust regression                               Number of obs     =        160
                                                F(  4,       155) =       2.51
                                                Prob > F          =     0.0444
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.000181	.0001097	-1.65	0.101	-.0003978	.0000357
vB11	.0062797	.0110796	0.57	0.572	-.0156067	.0281662
trained_quota100	.0606459	.086617	0.70	0.485	-.1104562	.2317481
expense_per_empl	.0291165	.0114247	2.55	0.012	.0065483	.0516848
_cons	33.31961	6.323078	5.27	0.000	20.82908	45.81014

Code.9

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1
  > & ATECO2007>=86 & ATECO2007<=88

    Huber iteration 1:  maximum difference in weights = .98989751
    Huber iteration 2:  maximum difference in weights = .96464813
    Huber iteration 3:  maximum difference in weights = .67338072
    Huber iteration 4:  maximum difference in weights = .35597922
    Huber iteration 5:  maximum difference in weights = .13800528
    Huber iteration 6:  maximum difference in weights = .02821953
Biweight iteration 7:  maximum difference in weights = .29507982
Biweight iteration 8:  maximum difference in weights = .08006199
Biweight iteration 9:  maximum difference in weights = .02336686
Biweight iteration 10:  maximum difference in weights = .00369831

Robust regression                               Number of obs     =        169
                                                F(  4,       164) =       3.16
                                                Prob > F          =     0.0157

------------------------------------------------------------------------------
productivity_h~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
          vB15_2 |   .0001368   .0000776     1.76   0.080    -.0000164       .00029
            vB11 |  -.0635958   .0502898    -1.26   0.208    -.1628948    .0357033
trained_quota100 |  -.3193108   .1380509    -2.31   0.022    -.5918972   -.0467245
expense_per_empl |   -.013024   .0142058    -0.92   0.361    -.0410738    .0150258
           _cons |   95.38446   11.29499     8.44   0.000     73.08211    117.6868
------------------------------------------------------------------------------
```

Code.10

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1
  > & ATECO2007>=1 & ATECO2007<=3
  insufficient observations
```

Code.11

```
2 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7==84
  no observations
  r(2000);
```

Code.12

```
2 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1
  > & ATECO2007>=97 & ATECO2007<=98
  no observations
  r(2000);
```

Code.13

```
3 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1
  > & ATECO2007==99
  no observations
  r(2000);
```

Code.14

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7==35

      Huber iteration 1:  maximum difference in weights = .96623486
      Huber iteration 2:  maximum difference in weights = .65650272
      Huber iteration 3:  maximum difference in weights = .36617294
      Huber iteration 4:  maximum difference in weights = .55488591
      Huber iteration 5:  maximum difference in weights = .27836755
      Huber iteration 6:  maximum difference in weights = .13084045
      Huber iteration 7:  maximum difference in weights = .10136044
      Huber iteration 8:  maximum difference in weights = .08350284
      Huber iteration 9:  maximum difference in weights = .09544127
      Huber iteration 10:  maximum difference in weights = .08950121
      Huber iteration 11:  maximum difference in weights = .01942578
    Biweight iteration 12:  maximum difference in weights = .27019057
    Biweight iteration 13:  maximum difference in weights = .17136424
    Biweight iteration 14:  maximum difference in weights = .03307563
    Biweight iteration 15:  maximum difference in weights = .01693458
    Biweight iteration 16:  maximum difference in weights = .01344518
    Biweight iteration 17:  maximum difference in weights = .00897625

    Robust regression                               Number of obs     =         46
                                                    F(  4,        41) =       1.17
                                                    Prob > F          =     0.3380

------------------------------------------------------------------------------
productivity_h~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
          vB15_2 |  -.0014505   .0032893    -0.44   0.662    -.0080935    .0051924
            vB11 |  -2.256373   1.717009    -1.31   0.196    -5.723943    1.211197
trained_quota100 |  -1.921115   2.313669    -0.83   0.411    -6.593664    2.751433
expense_per_empl |  -.0053592   .0453196    -0.12   0.906     -.096884    .0861656
           _cons |   586.5227   175.8994     3.33   0.002     231.2865    941.7588
------------------------------------------------------------------------------
```

Code.15

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7>=36 & ATECO2007<=39

     Huber iteration 1:  maximum difference in weights = .93121347
     Huber iteration 2:  maximum difference in weights = .29942222
     Huber iteration 3:  maximum difference in weights = .15011392
     Huber iteration 4:  maximum difference in weights = .04319852
    Biweight iteration 5:  maximum difference in weights = .29656102
    Biweight iteration 6:  maximum difference in weights = .05708074
    Biweight iteration 7:  maximum difference in weights = .01197577
    Biweight iteration 8:  maximum difference in weights = .00474542

    Robust regression                               Number of obs     =        197
                                                    F(  4,       192) =       0.85
                                                    Prob > F          =     0.4966

------------------------------------------------------------------------------
productivity_h~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
          vB15_2 |  -.0000107   .0001489    -0.07   0.943    -.0003043    .0002829
            vB11 |  -.1025385    .123495    -0.83   0.407    -.3461195    .1410425
trained_quota100 |   .4041906    .251944     1.60   0.110    -.0927428     .901124
expense_per_empl |   .0000751   .0153663     0.00   0.996    -.0302334    .0303835
           _cons |   115.7719   19.88943     5.82   0.000     76.54205    155.0017
------------------------------------------------------------------------------
```

Code.16

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7>=55 & ATECO2007<=56

     Huber iteration 1:  maximum difference in weights = .98480521
     Huber iteration 2:  maximum difference in weights = .97955376
     Huber iteration 3:  maximum difference in weights = .68663357
     Huber iteration 4:  maximum difference in weights = .58979917
     Huber iteration 5:  maximum difference in weights = .22508761
     Huber iteration 6:  maximum difference in weights = .13047098
     Huber iteration 7:  maximum difference in weights = .08682226
     Huber iteration 8:  maximum difference in weights = .06837959
     Huber iteration 9:  maximum difference in weights = .05106555
     Huber iteration 10:  maximum difference in weights = .01051717
    Biweight iteration 11:  maximum difference in weights = .29389913
    Biweight iteration 12:  maximum difference in weights = .11179743
    Biweight iteration 13:  maximum difference in weights = .05927209
    Biweight iteration 14:  maximum difference in weights = .00945252

    Robust regression                               Number of obs     =        132
                                                    F(  4,       127) =       0.50
                                                    Prob > F          =     0.7375

------------------------------------------------------------------------------
productivity_h~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
          vB15_2 |  -.0001675   .0002967    -0.56   0.573    -.0007546    .0004197
            vB11 |  -.0231814   .0264129    -0.88   0.382    -.0754478     .029085
trained_quota100 |   .0781528   .1355113     0.58   0.565    -.1899997    .3463053
expense_per_empl |   .0070038   .0109879     0.64   0.525    -.0147392    .0287469
           _cons |   62.23258   10.11875     6.15   0.000      42.2094    82.25576
------------------------------------------------------------------------------
```

Code.17

```
   Huber iteration 1:  maximum difference in weights = .9982355
   Huber iteration 2:  maximum difference in weights = .95165513
   Huber iteration 3:  maximum difference in weights = .69226534
   Huber iteration 4:  maximum difference in weights = .42125504
   Huber iteration 5:  maximum difference in weights = .21890569
   Huber iteration 6:  maximum difference in weights = .08355543
   Huber iteration 7:  maximum difference in weights = .01948962
Biweight iteration 8:  maximum difference in weights = .29291912
Biweight iteration 9:  maximum difference in weights = .11325506
Biweight iteration 10:  maximum difference in weights = .01792171
Biweight iteration 11:  maximum difference in weights = .00342948

Robust regression                               Number of obs     =        367
                                                F(  4,       362) =       0.25
                                                Prob > F          =     0.9075
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	7.48e-06	.0000201	0.37	0.710	-.000032	.0000469
vB11	.0023334	.0186106	0.13	0.900	-.0342651	.0389318
trained_quota100	-.0024972	.0591165	-0.04	0.966	-.1187522	.1137577
expense_per_empl	-.0001058	.0003706	-0.29	0.776	-.0008346	.0006231
_cons	56.50648	4.494816	12.57	0.000	47.66725	65.34571

Code.18

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7>=64 & ATECO2007<=66

   Huber iteration 1:  maximum difference in weights = .99738895
   Huber iteration 2:  maximum difference in weights = .84638243
   Huber iteration 3:  maximum difference in weights = .70750802
   Huber iteration 4:  maximum difference in weights = .42976246
   Huber iteration 5:  maximum difference in weights = .42107631
   Huber iteration 6:  maximum difference in weights = .52037147
   Huber iteration 7:  maximum difference in weights = .33777222
   Huber iteration 8:  maximum difference in weights = .33534208
   Huber iteration 9:  maximum difference in weights = .11883974
   Huber iteration 10:  maximum difference in weights = .07227935
   Huber iteration 11:  maximum difference in weights = .0326451
Biweight iteration 12:  maximum difference in weights = .27748239
Biweight iteration 13:  maximum difference in weights = .13346901
Biweight iteration 14:  maximum difference in weights = .04774171
Biweight iteration 15:  maximum difference in weights = .01678027
Biweight iteration 16:  maximum difference in weights = .00605742

Robust regression                               Number of obs     =        317
                                                F(  4,       312) =       2.22
                                                Prob > F          =     0.0667
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	.0002675	.0001322	2.02	0.044	7.37e-06	.0005276
vB11	-.0526879	.0421018	-1.25	0.212	-.1355273	.0301516
trained_quota100	-.3112275	.1980892	-1.57	0.117	-.7009872	.0785322
expense_per_empl	-.0024168	.0025647	-0.94	0.347	-.0074632	.0026295
_cons	127.2847	18.71326	6.80	0.000	90.46455	164.1049

Code.19

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7==68

   Huber iteration 1:  maximum difference in weights = .67385617
   Huber iteration 2:  maximum difference in weights = .23990509
   Huber iteration 3:  maximum difference in weights = .06991626
   Huber iteration 4:  maximum difference in weights = .08419004
   Huber iteration 5:  maximum difference in weights = .05761066
   Huber iteration 6:  maximum difference in weights = .01768255
Biweight iteration 7:  maximum difference in weights = .20572072
Biweight iteration 8:  maximum difference in weights = .05504918
Biweight iteration 9:  maximum difference in weights = .05820358
Biweight iteration 10:  maximum difference in weights = .04229183
Biweight iteration 11:  maximum difference in weights = .03486036
Biweight iteration 12:  maximum difference in weights = .02729434
Biweight iteration 13:  maximum difference in weights = .00575572

Robust regression                               Number of obs     =         14
                                                F(  4,         9) =       4.05
                                                Prob > F          =     0.0378
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	.0206714	.0209952	0.98	0.351	-.026823	.0681657
vB11	4.53876	9.372635	0.48	0.640	-16.66361	25.74113
trained_quota100	-.7246038	2.589906	-0.28	0.786	-6.583377	5.13417
expense_per_empl	-.3444093	.3536002	-0.97	0.356	-1.144308	.4554899
_cons	145.1294	168.2397	0.86	0.411	-235.4552	525.714

Code.20

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7>=69 & ATECO2007<=75

    Huber iteration 1:  maximum difference in weights = .99564525
    Huber iteration 2:  maximum difference in weights = .88448131
    Huber iteration 3:  maximum difference in weights = .59635048
    Huber iteration 4:  maximum difference in weights = .6213945
    Huber iteration 5:  maximum difference in weights = .20197201
    Huber iteration 6:  maximum difference in weights = .13021763
    Huber iteration 7:  maximum difference in weights = .12022657
    Huber iteration 8:  maximum difference in weights = .08620886
    Huber iteration 9:  maximum difference in weights = .0572955
    Huber iteration 10:  maximum difference in weights = .03104791
Biweight iteration 11:  maximum difference in weights = .29297046
Biweight iteration 12:  maximum difference in weights = .30708435
Biweight iteration 13:  maximum difference in weights = .15584766
Biweight iteration 14:  maximum difference in weights = .0862117
Biweight iteration 15:  maximum difference in weights = .03337364
Biweight iteration 16:  maximum difference in weights = .01359436
Biweight iteration 17:  maximum difference in weights = .00568619

Robust regression                               Number of obs     =        247
                                                F(  4,       242) =       1.08
                                                Prob > F          =     0.3653
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0000426	.0000378	-1.13	0.261	-.0001172	.0000319
vB11	.0562041	.0298264	1.88	0.061	-.0025484	.1149567
trained_quota100	-.0658968	.0744759	-0.88	0.377	-.2126005	.080807
expense_per_empl	.0049854	.0045728	1.09	0.277	-.0040223	.013993
_cons	63.69592	5.598878	11.38	0.000	52.66717	74.72468

Code.21

```
2 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7==85

    Huber iteration 1:  maximum difference in weights = .81773002
    Huber iteration 2:  maximum difference in weights = .22381939
    Huber iteration 3:  maximum difference in weights = .23825274
    Huber iteration 4:  maximum difference in weights = .11294709
    Huber iteration 5:  maximum difference in weights = .07770488
    Huber iteration 6:  maximum difference in weights = .0477076
Biweight iteration 7:  maximum difference in weights = .21851788
Biweight iteration 8:  maximum difference in weights = .25356935
Biweight iteration 9:  maximum difference in weights = .33772947
Biweight iteration 10:  maximum difference in weights = .31538203
Biweight iteration 11:  maximum difference in weights = .14374567
Biweight iteration 12:  maximum difference in weights = .02721693
Biweight iteration 13:  maximum difference in weights = .02579904
Biweight iteration 14:  maximum difference in weights = .00398962

Robust regression                               Number of obs     =         29
                                                F(  4,        24) =       0.21
                                                Prob > F          =     0.9281
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0004333	.0011226	-0.39	0.703	-.0027502	.0018837
vB11	.1052693	.1891828	0.56	0.583	-.2851848	.4957233
trained_quota100	.0773885	.2045586	0.38	0.709	-.3447998	.4995768
expense_per_empl	.0071413	.0113483	0.63	0.535	-.0162806	.0305631
_cons	34.91086	16.70952	2.09	0.047	.424105	69.39762

Code.22

```
1 . rreg productivity_hours vB15_2 vB11 trained_quota100 expense_per_empl if diversification>=1 & ATECO200
  > 7>=94 & ATECO2007<=96

    Huber iteration 1:  maximum difference in weights = .96174499
    Huber iteration 2:  maximum difference in weights = .6687464
    Huber iteration 3:  maximum difference in weights = .42919027
    Huber iteration 4:  maximum difference in weights = .46657237
    Huber iteration 5:  maximum difference in weights = .32230711
    Huber iteration 6:  maximum difference in weights = .07829163
    Huber iteration 7:  maximum difference in weights = .01392451
Biweight iteration 8:  maximum difference in weights = .29347465
Biweight iteration 9:  maximum difference in weights = .09569391
Biweight iteration 10:  maximum difference in weights = .01261724
Biweight iteration 11:  maximum difference in weights = .00532593

Robust regression                               Number of obs     =         93
                                                F(  4,        88) =       0.39
                                                Prob > F          =     0.8151
```

productivity_h~s	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vB15_2	-.0000564	.0002333	-0.24	0.810	-.0005201	.0004073
vB11	.0051183	.0432134	0.12	0.906	-.0807592	.0909958
trained_quota100	.0436727	.1119647	0.39	0.697	-.1788335	.2661789
expense_per_empl	.0031911	.0030088	1.06	0.292	-.0027882	.0091705
_cons	39.57851	9.063599	4.37	0.000	21.56652	57.59051